LIBRARY 178 MAY 19 1936

Vol. 33, No. 5 PSYCHOLOGICAL REVIEW PUBLICATIONS May, 1936

Psychological Bulletin

EDITED BY
JOHN A. McGEOCH, WESLEYAN UNIVERSITY

SAMUEL W. FERNBERGER, UNIV. OF PENNSYLVANIA *(J. Exper. Psychol.)*
WALTER S. HUNTER, CLARK UNIVERSITY *(Psychol. Index)*
HERBERT S. LANGFELD, PRINCETON UNIV. *(Psychol. Rev.)*
JOHN F. DASHIELL, UNIV. OF NORTH CAROLINA *(Psychol. Monog.)*

WITH THE CO-OPERATION OF
G. W. ALLPORT, HARVARD UNIVERSITY; J. E. ANDERSON, UNIVERSITY OF MINNESOTA; W. T. HERON, UNIVERSITY OF MINNESOTA; J. T. METCALF, UNIVERSITY OF VERMONT; R. PINTNER, COLUMBIA UNIVERSITY.

CONTENTS

PUBLISHED MONTHLY (EXCEPT AUGUST AND SEPTEMBER)
FOR THE AMERICAN PSYCHOLOGICAL ASSOCIATION
BY THE PSYCHOLOGICAL REVIEW COMPANY
PRINCETON, N. J.

Entered as second-class matter at the post-office at Princeton, N. J., with an additional entry at Albany, N. Y.

Publications of the American Psychological Association

EDITED BY

S. W. FERNBERGER, University of Pennsylvania (*J. Exper. Psychol.*)
WALTER S. HUNTER, Clark University (*Index and Abstracts*)
HENRY T. MOORE, Skidmore College (*J. Abn. and Soc. Psychol.*)
HERBERT S. LANGFELD, Princeton University (*Review*)
JOHN A. McGEOCH, Wesleyan University (*Bulletin*)
JOHN F. DASHIELL, University of North Carolina (*Monographs*)

HERBERT S. LANGFELD, Business Editor

PSYCHOLOGICAL REVIEW

containing original contributions only, appears bi-monthly, January, March, May, July, September, and November, the six numbers comprising a volume of about 540 pages.

PSYCHOLOGICAL BULLETIN

containing critical reviews of books and articles, psychological news and notes, university notices, and announcements, appears monthly (10 numbers), the annual volume comprising about 720 pages. Special issues of the Bulletin consist of general reviews of recent work in some department of psychology.

JOURNAL OF EXPERIMENTAL PSYCHOLOGY

containing original contributions of an experimental character, appears bi-monthly, February, April, June, August, October, and December, the six numbers comprising a volume of about 900 pages.

PSYCHOLOGICAL INDEX

is a compendious bibliography of books, monographs, and articles upon psychological and cognate topics that have appeared during the year. The Index is issued annually in June, and may be subscribed for in connection with the periodicals above, or purchased separately.

PSYCHOLOGICAL ABSTRACTS

appears monthly, the twelve numbers and an index supplement making a volume of about 700 pages. The journal is devoted to the publication of non-critical abstracts of the world's literature in psychology and closely related subjects.

PSYCHOLOGICAL MONOGRAPHS

consist of longer researches or treatises or collections of laboratory studies which it is important to publish promptly and as units. The price of single numbers varies according to their size. The Monographs appear at irregular intervals and are gathered into volumes of about 500 pages.

JOURNAL OF ABNORMAL AND SOCIAL PSYCHOLOGY

appears quarterly, June, September, December, March, the four numbers comprising a volume of 448 pages. The journal contains original contributions in the field of abnormal and social psychology, reviews, notes and news.

ANNUAL SUBSCRIPTION RATES

Review: $5.50 (Foreign, $5.75). Index: $4.00 per volume.
Journal: $7.00 (Foreign, $7.25). Monographs: $6.00 per volume (Foreign, $6.30).
Bulletin: $6.00 (Foreign, $6.25). Abstracts: $6.00 (Foreign, $6.25).
Abnormal and Social: $5.00 (Foreign, $5.25). Single copies $1.50.
Current numbers: Journal, $1.25; Review, $1.00; Abstracts, 75c; Bulletin, 60c.

COMBINATION RATES

Review and Bulletin: $10.00 (Foreign, $10.50).
Review and J. Exper.: $11.00 (Foreign, $11.50).
Bulletin and J. Exper.: $12.00 (Foreign, $12.50).
Review, Bulletin, and J. Exper.: $16.00 (Foreign, $16.75).
Review, Bulletin, J. Exper., and Index: $19.00 (Foreign $19.75).

Subscriptions, orders, and business communications should be sent to the

PSYCHOLOGICAL REVIEW COMPANY

PRINCETON, N. J.

Vol. 33, No. 5 May, 1936

THE
PSYCHOLOGICAL BULLETIN

THE METHODOLOGY OF EXPERIMENTAL STUDIES OF HUMAN LEARNING AND RETENTION: I. THE FUNCTIONS OF A METHODOLOGY AND THE AVAILABLE CRITERIA FOR EVALUATING DIFFERENT EXPERIMENTAL METHODS

BY ARTHUR W. MELTON
University of Missouri

The present review is the first of three that will be concerned with the methods, materials, and measures used in experimental investigations of human learning and retention. Succeeding reviews will undertake to summarize and evaluate the specific methods, materials, and measures used in studies of ideational or verbal learning and retention (memory) and in the investigation of the learning and retention of motor habits, including the maze. Since any attempt to evaluate different methods [1] in terms of their adequacy for experimental studies of the different types of human learning requires the acceptance of certain assumptions regarding the need for a precise methodology and the acceptance of certain criteria in terms of which the different methods commonly used may be compared, this first paper has been devoted entirely to the explication of those assumptions and criteria.

The study of learning and retention in the human subject has been one of the most active fields in experimental psychology since the pioneer work of Ebbinghaus on memory in 1885 (**28**). These fifty years have witnessed not only a considerable increase in the

[1] *Method* is used as the generic term for all aspects of the experimental situation, and also in referring specifically to the aspects of the situation other than the material or task learned and the measures of learning and retention used.

knowledge of the factors which determine the degree of learning and retention of verbal materials, but also an extension of the field of investigation to motor habits or skills in the work of Bryan and Harter on telegraphy (17), of Book on typewriting (9), and in the extremely significant application by Hicks and Carr (42) and Perrin (91) of the maze task to the study of human learning. Concurrent with these advances and extensions of the field of investigation, there has necessarily been a marked increase in the complexity of the methodology of the field. Each new type of learning or retention subjected to investigation required significant changes in the techniques used. For example, there arose special methods for studying serial rote learning, the formation of single associative connections, and trial and error learning in the formation of both verbal and motor habits, as well as methods for determining the degree of retention in terms of recall, relearning or saving, recognition, and "reconstruction." Likewise, significant advances have been made in the design of experimental studies, such that the investigators may specify more exactly the conditions which are operative at the time of learning and retention, and may, therefore, have greater assurance that unwanted variable factors, such as the practice effect from one session of learning to the next, have not obscured or invalidated their experimental findings. As a consequence of these various directions of improvement and elaboration in the experimental techniques, there is at the present time a great variety of materials, measures, and techniques of experimental control that are available for use in experiments on human learning and retention.

This diversity of methods undoubtedly represents the continuing interest of psychologists in the methodological problems presented by the field of study, but the disturbing feature of the methodology is that no set of principles or practices finds universal acceptance. Although many investigators use nonsense syllables in the study of serial verbal learning, many others use lists of numbers, words, or combinations of numbers and letters, even though their experimental studies have not been framed with reference to the characteristics of the material used. Similarly, in maze studies different investigators use multiple-T mazes, a Warden U-maze, or mazes of their own design in the study of the effect of a particular variable, even though the type of maze alley or the length of the maze is not considered as an experimental variable. The case is the same with many of the constant conditions other than the material learned. The time interval during which individual units in a list are presented varies

in recent experiments from 0.75 seconds to 4.00 seconds, although none of the experimenters considered the time interval as an essential condition for the study of the phenomena of major concern. Still another example of this fortuitous variation of a basic constant from experiment to experiment is found when one examines the criterion used by different investigators to define "complete mastery"; some have used the criterion of one errorless trial, others have used two errorless trials, and still others demand the fulfilment of a more stringent criterion. As a result, one investigator's "complete learning" is another's "overlearning." Even the formulae used to derive some of the basic measures of learning and retention, such as the saving score (**43**), are not universal in form, but vary from experimenter to experimenter.

These are merely illustrative samples of the variation permitted by the existing methodology in the conditions of experiments. Obviously, if there are specific rules or principles of procedure that are valid and that should be adhered to except when the experimental problem demands a modification, they have only limited recognition. The clue to this situation is given by the dependence of our methodology on rationalistic criteria for the evaluation of methods, or on tradition, rather than on specific experimental determinations of the most adequate materials and techniques for the study of learning in the human subject. Thus, Ebbinghaus invented the nonsense syllable because he *believed* that it would be less variable than meaningful materials with respect to the difficulty of learning, and many investigators have accepted the nonsense syllable as an adequate material because it was used by Ebbinghaus. But tradition and untested beliefs regarding the amount of variable error which will result from the use of certain methods are not sufficiently potent to effect a standardization of methods of research. Why should one adhere to the use of the nonsense syllable merely because Ebbinghaus *believed* that it constituted the most adequate experimental material, especially when later investigators have been free with statements of their contradictory *beliefs?* Unless one answers that any reasonably satisfactory technique should be generally accepted because the standardization of the methods of research is necessary for the systematization of the results of different experiments, the answer must be that standardization cannot be expected when there are no data regarding the relative adequacy of the several "traditional" techniques offered to those who perform the experiments.

The unquestionable need for an experimental approach to the

problems involved in the selection of methods of research to be used in studying human learning, and the correlative need for objective measures or indices in terms of which the comparison of methods may be effected, lends particular significance to the article published by Spearman in 1910 on the "reliability coefficient" and its use in evaluating experimental materials (**109**). Although this paper led almost immediately to extensive use of the reliability coefficient as an index in terms of which mental and educational tests were evaluated, its significance was not generally recognized by students of human learning until the report of Hunter in 1922 on the reliability of the maze as an instrument for the study of human and animal learning (**49**). His challenging conclusion that the maze was unsatisfactory as a learning task and that experimental results obtained with it were questionable led to the controversy with Carr (**18, 50**) which served not only to clarify the function of measurements of reliability in the development of adequate methods of research, but also stimulated a number of investigators to use the reliability coefficient in the comparison of different learning materials and methods. As a consequence, mazes commonly used in the study of human and animal learning were found to yield measurements that differed markedly in reliability. When the different measures of learning, such as trials, time, and errors, were compared, they too were found to be unequally reliable. The materials used in the study of ideational or verbal learning were eventually subjected to similar experimental and statistical comparisons, and again the value of the empirical approach was illustrated. For example, nonsense syllables were found to vary greatly in meaningfulness (**37, 47, 61**), and at least one investigator (**26**) concluded that Ebbinghaus was in error in believing that lists of nonsense syllables varied less in difficulty than lists of words. In short, the application of the single criterion of reliability to the evaluation of traditional techniques resulted in sufficiently provocative conclusions to justify the contention that a methodology based on *a priori* or casually empirical grounds should be replaced by a methodology based on experiment.

When the advantages and disadvantages of various materials and procedures in the study of human learning can be removed from the realm of individual conjecture and either placed on a solid empirical basis or defended in terms of certain clearly stated postulates, we may expect an increase in the validity and reliability of particular experimental results, and perhaps a standardization of experimental procedures. However, it is an error to believe either that the prob-

lems involved in the use of the reliability of a method as an index of its value have been solved, or that the reliability of the method is the sole criterion of its value. It is nearer the truth to believe that the explicit and implicit disagreement among students of learning regarding the most adequate materials and methods for use in research has merely been replaced by explicit and implicit disagreement regarding the objective measures and criteria to be employed in evaluating the different aspects of their methodology. There is no general agreement regarding either the meaning of the reliability coefficient or the methods to be used in determining the coefficient when two or more materials or methods are to be compared in terms of it (p. 364 ff.), and these difficulties are accentuated when one attempts to apply the criterion in the evaluation of methods used to study human learning. Most of the discussions of the use of the reliability coefficient have been specific to the experimental situations involved in the study of animal learning, rather than human learning, and to maze learning, rather than to the learning of verbal materials. It is, therefore, highly probable that the various ways of determining the reliability of methods used with human subjects and in the study of ideational or verbal learning must differ from the ways considered adequate for studies involving the white rat and mazes. As a further complication, there is the problem of evaluating other proposed indices of reliability, such as the measures of absolute and relative variability used by Davis (**26**), Sauer (**104**), McGeoch (**77**), and Stroud, Lehman, and McCue (**115**), and the critical ratio (difference/$\sigma_{difference}$) as used by Maurer and Carr (**75**). Unanimity of opinion regarding the significance and validity of these indices must be attained before conclusions based on them are permitted to determine the methods used in specific research studies.

Furthermore, there is the question of other criteria besides reliability for use in evaluating materials and methods. Most investigators would probably agree that some form of meaningless verbal material must have a place in the methodology of human learning, even though it is shown to yield rather unreliable measurements, because it is needed in studies which propose to examine the relationships between the meaningfulness of the material learned and some phenomenon such as "reminiscence." It might be contended that the method has *prima facie* validity. However, it is conceivable that in many instances this attribute of a material or method may likewise be subjected to measurement, and that the relative validity of different materials or methods may be determined experimentally.

Such has been the suggestion made by Commins, McNemar, and Stone (21) as a preface to their study of the community of function between the abilities measured by mazes, problem boxes, and discrimination boxes. Still other criteria are suggested when one attempts to evaluate different techniques for the control of variable or progressive errors, such as practice and fatigue, which are wont to occur in experiments. As an example, the methods commonly used to control practice effects when the same subjects are run through the different conditions of an experiment may be evaluated in terms of the extent to which the data obtained by the different methods are susceptible to accurate statistical analyses. Some methods yield data that fail to satisfy the assumptions involved in the use of the common statistical formulae for the estimation of the dependability of the obtained experimental differences, and other methods yield data that satisfy those assumptions. These and other criteria must obviously be examined carefully before essaying a critical evaluation of existing methods for studying human learning.

The Functions of a Methodology

The aggregation of methods of experimental control and measurement which is the methodology of a field of research has two distinguishable, but closely related, functions that have been implied in the preceding discussion. The first and most obvious function is that of listing the factors which must be measured or controlled in experiments in the field, and of stating the most adequate methods for measuring or controlling those factors. In this way, an adequate methodology serves to increase the *validity* and *reliability* of experimental conclusions. The *validity* of the interpretation of the experimental result is increased because the investigator becomes more certain that the difference between the performance of his subjects under the experimental and control conditions is attributable to the intentionally varied factor, and not to that factor plus some uncontrolled factor, such as practice, fatigue, misinterpreted instructions, unequal difficulty of the materials learned, etc. Furthermore, the *validity* of the interpretation is increased by the fact that the investigator can specify more exactly the particular conditions which were operative in his experiment, quite apart from the question of constant experimental errors.

A significant example of this function of a methodology is given by the recent increase in the exactness with which investigators of the phenomena of rote learning have been able to specify the conditions operative in their experiments

as a consequence of Glaze's (37) calibration of nonsense syllables in terms of their association-value. Whereas for many years investigators merely asserted that their records were for the learning of nonsense syllables, it is now possible to state, with a degree of accuracy which is limited only by the accuracy of Glaze's determinations, the degree of meaningfulness of the "nonsense syllables" employed, and some of the uncertainty regarding the justification of the investigator's conclusion that "Condition *X* is effective with *meaningless* materials" may be removed.

As a second consequence of this function, the *reliability* of the experimental result is increased because the most adequate methods of control are those which yield the minimal amounts of variable error. That is, assuming an adequate control of potential constant experimental errors by the conversion of many of them into variable errors with means of zero, the application of the most adequate methods of control increases the dependability of the obtained experimental difference or lack of difference by reducing the variability of the obtained individual measurements.

This first function of a methodology has to do, therefore, chiefly with the economy of research efforts. The lack of an adequate methodology cannot prevent the progress of experimental analysis, but it may delay progress toward confident generalization either by permitting false interpretations or by rendering the attainment of dependable experimental determinations of the effect of certain factors too expensive of time and labor. However, an invalid conclusion that rests on an inadequate methodology will in time be corrected by repetitions of the experiment, and the unreliable conclusion may be converted into a reliable conclusion by the accumulation of more data. Presumably, the importance of methodology when viewed with reference to this practical function will not be questioned.

It is apparent that the first function has to do with the acceptability of the conclusions of isolated experiments. A second function, and one which appears to the reviewer to be of even greater significance, is that of promoting the standardization of the experimental conditions and measures used in studies of human learning. A complete and adequate methodology should not only increase the validity and reliability of the isolated experiment, but it should also increase the opportunity for a true systematization of the results of many experiments. This systematization cannot be precise when the basic constant conditions vary from experiment to experiment, and when there are no available means of estimating the consequences to be expected from such variations.

One has only to attempt to generalize the findings of different experimenters on almost any problem in the field of learning and

retention in order to become convinced of the advantages to be derived from a standardization of techniques—not only the materials or tasks used, but also the techniques of presentation, the techniques for controlling practice effects and other intra-individual consequences of repeated learning, and the techniques for measuring the amount learned, learning time, or amount retained. One experimenter studies the distribution of practice in the learning of nonsense syllables, and uses, among other conditions, a three-second presentation of each unit, and a time-limit method for determining the degree of learning; whereas, another experimenter studies the "same" problem with lists of three-place numbers, with a two-second presentation of each unit, and with a work-limit determination of the degree of learning.

The philosophy behind such multiple variation in the experimental conditions in studies which are presumed to be directed toward the determination of the effects of the same variable, *e.g.* the distribution of practice, has been appropriately labelled by Carr (20) as "The Quest for Constants." A similar criticism has been stated by Johnson (56). There appears to be the belief that the laws of learning and retention are to a great extent independent of the experimental operations performed in the discovery and verification of them. Carr's and Johnson's citations are ample evidence against the validity of this assumption, and Robinson (100, p. 129) has formalized this reaction against the quest for constants in learning by defining *the* law of association as $A=f(x, y, z \ldots)$, where A is associative strength, and x, y, z, etc., are such factors as time interval, frequency of repetition, the state of other existing associative connections, etc. The success of investigators in carrying this conception of the association theory into the laboratory, and in making the theory productive of a greater understanding of the phenomena of learning, depends upon the accuracy of the quantitative evaluations (calibrations) of the factors involved in the determination of associative strength. This in turn rests on the acceptance of some set of conditions as a starting point for the investigation of each variable in its relation to every other variable. In short, standardization of conditions and of general methodology is propaedeutic to systematization of experimental facts.

Those who prefer to emphasize the deductive process, rather than the inductive process, in the development of a scientific systematization of the phenomena of learning are no less in need of a body of experimental facts that have been determined under con-

stant conditions. Hull's recent outline of a miniature scientific theoretical system for the explanation of certain phenomena of serial learning makes this need apparent (**48**). If the experimental test of a deduction is to determine the validity of the "postulates," rather than the fallibility of the specific deduction formulated, the deduction must be stated in quantitative terms and must refer to a specific set of conditions. The deduction of the algebraic resultant of the interaction of two or more factors is *ipso facto* fallible when the effect of each factor is stated in the semi-quantitative form of "moreness" or "lessness," and when the specific experimental operations involved in the determination of the effects of each variable in isolation (the "postulates") have not been constant for all the variables involved.

Standardization, in the sense in which this term is used here, does not imply a restriction of all experimental studies to a certain set of conditions. For example, it is not a true inference that all studies should be conducted with multiple-T mazes, with nonsense syllables, with presentation intervals of 2 seconds, etc. Standardization means the acceptance of a standard set of conditions to which all deviations from the standard conditions may be referred. In this sense a standard set of conditions and measures for use in learning experiments would be to the systematization of the facts of learning as the vacuum and the *c.g.s.* system is to the systematization of the facts of physics. Experiments which differed radically from each other in the conditions considered as constant would still be performed, but with the difference that there could be some confidence in the attempts to integrate the results of these experiments.

This may seem to be the setting of a goal for which psychology is not prepared, as Köhler maintains (**60**), or a goal that psychology can never approximate, as Bartlett maintains (**8**). According to the latter, psychologists should abandon their attempts to model their experiments and systematizations after the manner of the physical sciences and accept the clinical approach. In particular, he stresses the sterility and confusion of experimental work on human learning, and criticizes Ebbinghaus for setting the mode with respect to the standardization of techniques. One answer to this objection is that intelligible and useful systematizations of the relationship between a limited number of variables have been obtained by investigators who have performed a series of experiments in which the same basic conditions have been used as a reference point. In fact, the possibility of systematization is revealed by any investigation or series of

investigations that involves the systematic variation of more than two factors. Such an investigation or series of investigations may, in fact, qualify as a miniature system. As examples of such miniature systems, and the significant contributions to knowledge made by them, may be cited the work of Ebbinghaus on the factors that determine the degree of retention (**28**), the work of Luh on the conditions of retention of verbal material (**68**), the work of Carr and his students on guidance in learning (**19**), the work of Robinson and Heron (**102**) and Robinson and Darrow (**101**) on the relation between the amount of rote material learned, the degree of retention, and the degree of retroactive inhibition, the work of McGeoch on the factors determining the degree of retroactive inhibition (**76**, **79**, **80**, **81**), and the work of Skinner (107) on the factors determining the conditioned response in the rat. In the opinion of the writer, the confusion and sterility of which Bartlett speaks is not so much a consequence of the emulation of the basic methodological principles of the physical sciences as it is a consequence of the failure to worship incessantly at their shrine of standardization. To date, the standardization of conditions and methods for the study of learning has been limited to single systematic experiments, the work of individual investigators, or, at most, the research in single laboratories. To say that students of learning have modeled their research efforts after the manner of physics is to imply that each physicist makes his meter stick suit the size of his laboratory cupboard, uses his personal watch as an accurate indicator of sidereal time, and neglects to consider the atmospheric pressure at his particular geographic location. Psychologists have yet to achieve inter-laboratory standardization.

In setting inter-laboratory and inter-experiment standardization of conditions and methods as an important function of a precise methodology there is, however, one assumption that may be subject to question, namely, the assumption that *necessary* differences between the experiments performed in different laboratories and by different experimenters do not preclude the possibility of obtaining comparable experimental measurements. Obviously, differences between the human material with which different experimenters work may be taken into account by proper normative studies; differences in the laboratory environment may be minimized; and other factors in the situation, such as the material used, the mode of presentation, the instructions, etc., may be held constant. But the experimenter is the one factor in every study of human learning that cannot be held constant, and his effect on the performance of the subjects may be

sufficient to preclude complete standardization and systematization. In studies involving rats this source of error has been minimized by the use of automatic devices for handling the rats (**124, 130**), but in the case of studies of human learning the answer to the question regarding the effect of the experimenter must be given by direct comparisons of the results obtained by different experimenters when the experimenters have been equally trained. Several studies of this type have been made recently and the evidence favors the view that the "personal equation" of the experimenter is not of sufficient importance to preclude inter-experiment systematization if the experimenters are trained.

Barr (**7**) had 4 experimenters work with groups of 16 subjects. Each subject learned 2 lists of 10 nonsense syllables and 2 lists of 15 pairs of words. One list of each material was learned during the intermittent presentation of a 150-watt light and a loud buzzer, and the other list of each material was learned in the absence of the "distracting" stimuli. All 4 experimenters obtained similar results, and the differences between the mean scores made by the different groups of subjects on the same condition were explainable as sampling errors. McGeoch (**82**) obtained similar results in a much more complicated experiment. In this study 3 different experimenters worked with groups of 20 subjects in a study of retroactive inhibition which included 5 experimental conditions, and 1 of the experimenters ran 2 complete groups of 20 subjects. The agreement between the measures obtained from the 4 groups of subjects was remarkable; the measures obtained by the different experimenters differed by an amount no greater than that to be expected by chance, and the difference between the 2 groups run by the same experimenter was as great as the differences between the groups run by different experimenters. When experimenters are carefully trained with respect to their methods of approach to the subjects and in the details of the recording of responses, as were McGeoch's, their results are quantitatively comparable. It seems, therefore, that the systematization of experimental results obtained by different experimenters is a possibility which awaits only the standardization of the methods of experimentation.

If it is granted that a precise systematization of the experimental facts of human learning is feasible, and awaits only the standardization and calibration of experimental conditions, controls, and measures, then an adequate and complete methodology has the second function which we have ascribed to it. Although the conditions to be accepted as standard may be determined by *fiat*, and many of the conditions and controls must be standardized in this way until data regarding their relative adequacy according to other criteria are available, it is plausible to accept those specific conditions which research has shown to yield the most valid and reliable

results as the standard conditions to which all deviations are referred. Therefore, the immediate problem set by both functions of an adequate methodology reduces to that of discovering the most valid and reliable methods of control and measurement in studies of human learning. The most valid and reliable methods will be the ones selected as standard and will be the most productive of interpretable results in isolated experiments.

Criteria for Evaluating Experimental Methods

The purpose of a set of criteria is to enable the investigator to select those methods and conditions of experimentation that will yield measurements in terms of which a dependable and true answer to a specific question may be determined. The specific criteria may therefore be placed in one of two groups, those that refer to the *validity* of the measurements, and those that refer to the *reliability* or *dependability* of the measurements. In a complete methodology it should be unnecessary to have more than these two major categories of criteria. However, the methodology of research on human learning is far from complete; there are many aspects of the experimental situation that may be treated in several different ways, and yet there are at present insufficient data for even a tentative conclusion regarding the reliability or validity of the several alternatives. In such cases the appeal must be to a third criterion, namely, conformity with the conditions of other experiments that have major systematic significance in the field. It is obvious that this third criterion is important chiefly as a means for standardizing procedures and thus for promoting the systematization of experimental results.

The distinction between the concept of validity and the concept of reliability is basic to the discussions that follow. It is, of course, apparent that every individual measurement is "true" in the sense that it represents accurately the combined effect of all the conditions of the subject, the experimenter, and the environment at the moment the measurement is made. Thus, if a subject requires 15 trials to master a ten-unit list of nonsense syllables at a certain hour on a certain day while under the observation of a certain experimenter, the measurement is a perfectly valid and reliable indicator of the combined effect of the basic ability of the subject, plus his momentary state, plus the material learned, plus the distractions that may have occurred, plus the errors the experimenter may have made in recording the subject's progress, etc. This conclusion is demanded by a thoroughgoing determinism, but it is mere sophistry when considered

within the frame of reference delineated by the aims of science—analysis, systematization, and prediction. The validity and reliability of the measurement is perfect only for the unique combination of circumstances prevailing during the experimental observation, and science cannot deal with events that are unique. Science can deal only with recurring aspects of events.

The single measurement is a fallible indicator of the average conditions prevailing throughout a series of measurements, since it is partly, if not wholly, determined by variable factors. *In so far as the measurement is not an exact indicator of the average conditions prevailing throughout the series of measurements it lacks reliability. In so far as it is not an exact indicator of factors that the investigator identifies, but is, rather, an indicator of the effect of unidentified or erroneously identified factors, the measurement lacks validity.*

If, for example, each of 5 observations are determined by factors that have the following values, or weights, at the moment the observation is made:

1st	Measurement	4A	8B	7C	2D	(0E)	(0F)
2nd	Measurement	4A	7B	9C	1D	(0E)	(0F)
3rd	Measurement	4A	9B	5C	(0D)	1E	1F
4th	Measurement	4A	5B	6C	4D	1E	(0F)
5th	Measurement	4A	6B	8C	3D	(0E)	(0F),

the average of the measurements will represent the average condition 4A, 7B, 7C, 2D, .4E, .2F, after the proper positive and negative signs have been given to each factor. The measurement obtained during any single observation will be a fallible indicator of this complex of conditions, and will thus lack reliability. If, furthermore, factors *A*, *B*, and *C* are identified correctly and factors *E* and *F* are unidentified or incorrectly identified, the interpretation of the single measurement or the average of the single measurements will lack validity, because the obtained measurement represents more than, or something other than, the factors it is thought to represent.

This example reveals the close relationship between validity and reliability. The reliability of a single measurement depends not at all on the identification of the factors, and their weights, but merely refers to the deviation of the individual measurement from the mean of a large number of measurements obtained under the "same" conditions. This deviation is obviously caused by a variation in the weights given to the various factors in the complex or by the intrusion of sporadic factors, and the reliability of the measurement increases as the amount of variation in the conditions of measurement decreases. Since the conditions of measurement may be highly constant and yet the investigator be unable to specify with exactitude the nature of the factors measured, the reliability of measurements may

be high and yet the validity may be either high or low. On the other hand, when the reliability of measurements is low, the weights given the various factors have varied greatly throughout the series of measurements and the sporadic factors have been abundant. Therefore, it is impossible for the investigator to identify the factors operating to determine the measurements, and the *validity* of the measurements must be low. In short, as the reliability of measurements approximates zero as a limit, the complex of factors determining each single measurement becomes more nearly unique, and the validity of the measurement obtained in the unique situation is zero when we consider the measurement in terms of its usefulness for analysis and prediction.

The concepts of validity and reliability may be applied either in the evaluation of the results of a particular experiment or in evaluating the methods, materials, and measures used in determining those results. It is the latter use in which we are most interested at the present time, but this interest is occasioned by the intimate relationship between the validity and reliability of the experimental methods used in the determination of particular experimental results and the consequent validity and reliability of the experimental results themselves. As will be pointed out later (p. 334 ff.), the possibility of determining a reliable experimental result, *i.e.* a result that cannot be interpreted as an effect of chance factors, increases as the reliability of the methods used increases. Likewise, the validity of the interpretation of the particular experimental result increases as the validity, *i.e.* interpretability, of the measurements obtained by the method used increases. However, it is well to keep in mind the fact that every criterion of the reliability or validity of some aspect of the experimental method employed in the study of human learning is valid only in so far as it enables the selection of a procedure, material, or measure that will give the most dependable and interpretable results when used in a specific experimental study.

I. VALIDITY AS A CRITERION

The essence of the concept of validity as applied in the evaluation of the methods, materials, and measures used in the study of human learning is the reference to the adequacy of the identification of the factors that determine the value of a measurement or the mean value of a series of measurements. In so far as the interpretation of an experimental result is directly conditioned by the operations performed in producing and measuring a phenomenon, the definition of

the phenomenon must be given in terms of those operations. This principle of operational definition is implicitly accepted by most experimentalists, and has recently been explicitly adapted to psychological concepts by Stevens (**111, 112**) and McGeoch (**83**). One of the important consequences of the operational definition of scientific concepts is the extraordinary emphasis placed on problems of method. In particular, the emphasis is placed on questions regarding the validity of methodological practices, because operationalism places a premium on precise identification of the operations performed in describing a phenomenon. In the case of human learning these operations are the procedures, methods of control, materials, and measures used by the investigator. Furthermore, a thoroughgoing operationalism must go beyond the particularized operations, *i.e.* "conditions," employed in a single experiment and seek to give these operations membership in broader classifications. This process of classification is a source of error in the interpretation of experiments, and introduces the problem of validity and the need for a criterion of validity.

Types of Invalidation. The interpretation of a particular experimental result, *e.g.* the discovery that condition X yields a mean trial score of 10 and that condition Y yields a mean trial score of 15, obviously depends on the completeness and accuracy of the investigator's identification of the factors operative in condition X and in condition Y, and, therefore, of the difference between the factors operative in conditions X and Y. In general, it is helpful to distinguish between the accuracy with which the investigator defines the basic conditions operative in both conditions X and Y and the accuracy with which he defines the *difference* between conditions X and Y.

As an illustration of the usefulness of this distinction in considering the validity of methods of experimentation, consider an experiment in which the "control" subjects learn a stylus maze under "normal" conditions and the "experimental" subjects learn the maze while grasping a dynamometer. The experimenter has, as far as he knows, held all factors constant except the experimental factor, the muscular tension resulting from the grasping of the dynamometer, and he discovers that the "experimental" subjects learn the maze in fewer trials than the "control" subjects. Even though the obtained difference is statistically reliable, an interpretation of this result may be invalid in either one of two ways.

If the investigator concludes that increased muscular tension, or the grasping of the dynamometer, increases the speed of learning, this interpretation may be invalid because the grasping of the dynamometer was not the only constant difference between the situations determining the performance of the

"experimental" and "control" subjects. Although the investigator has taken a number of precautions to insure the equivalence of the basic conditions, there may have been some other constant difference between the conditions. If this occurs, it is an invalidation of the experimental method used.

On the other hand, the investigator's interpretation of his experiment may be invalid even though the increased muscular tension was the *only* difference between the 2 conditions, because the investigator is inaccurate in his definition of the basic conditions involved in the experiment, *i.e.* because he is inaccurate in his identification of the class of conditions, or operations, to which his particular operations belong. Thus, he may conclude that muscular tension increases the speed of formation of motor habits, that muscular tension increases the speed of formation of a stylus-maze habit, or that muscular tension increases the speed of "learning," and each of these statements may be questioned on the grounds that his material (the particular stylus maze used), his measures (*e.g.* trials alone), or his general procedure (*e.g.* the nature of the instructions, the amount of prior knowledge of the maze possessed by the subjects), were not representative of the class of operations that define "learning," "motor learning," or "stylus maze learning." This failure to identify correctly the class of operations to which the particular operations used in the experiment belong is the second source of invalidation.

Thus, an error in the interpretation of an experimental result may be attributed either to the occurrence of an "experimental error" or to the failure of the experimental operations or methods to represent adequately the operations or methods used to define the phenomenon indicated in the interpretation. In either case it is apparently a problem that centers about the experimental method, and it is legitimate to attribute the validity or lack of validity of an interpretation to the validity or lack of validity of the experimental method. Accordingly, there is a need for criteria in terms of which the validity of experimental methods may be evaluated and compared such that both sources of error in interpretations may be avoided.

A. Representativeness as a Criterion of the Validity of an Experimental Method

It is clear that the second source of error in the interpretation of experimental results reduces to the question: Are the methods, materials, and measures employed in the experiment representative of the class of operations that defines the phenomenon specified in the interpretation of the experiment? That is, does the experimental method yield measurements of "learning," "motor learning," "stylus maze learning," "immediate memory," etc., as the investigator has assumed in his interpretation? If the methods, materials, and measures used in the study of human learning are not equally

representative of a general learning ability, or if they are not equally representative of special forms of learning that may be designated, it is apparent that there is a real need for *the evaluation of these various aspects of experimental methods in terms of the degree to which they represent the operations common in all studies of learning or the operations common in restricted types of learning.*

1. *Specificity of the Learning Measured by Different Methods, Materials, and Measures.* The evidence from experimental studies of the community of function in different methods of studying learning unquestionably favors the view that such a criterion of the validity of a specific method is needed. The evidence takes two forms. In the first place, it is well known that the correlation between measures of "learning" obtained with different experimental methods that purport to represent "learning" is often very low, and sometimes so close to zero in a series of experiments that there is legitimate reason to believe that the methods represent entirely distinct functions. Many very low coefficients are cited in Anastasi's (2) summary of the studies of the relationships between different measures of "memory," and the average intercorrelation of the 8 memory tests used by her was only 0.29 when uncorrected for attenuation. The correction for attenuation due to unreliability of the single tests increased the average r to 0.40.

Hall (38) has recently summarized some of the studies of the intercorrelation of measures of human learning obtained with different methods and materials, and has presented important new data on the correlation between scores made by subjects in mastering a punch-board maze, a stylus maze, a Peterson Rational Learning Problem, and a list of nonsense syllables. In the summary of 84 r's obtained in previous studies (in which the variables correlated were measures of improvement in color naming, cancellation, opposites, addition, mental multiplication, typewriting, digit-symbol substitution, Turkish-English vocabulary, code learning, rational learning, checker puzzle, stylus maze, inverted writing, number completion, tapping, and word building) Hall found the median positive coefficient to be only 0.25. Only one-third of the total number of coefficients was significantly different from zero. These coefficients were not corrected for attenuation due to the unreliability of measurements, and there may have been systematic factors in the experiments that operated to attenuate them. Nevertheless, the conclusion that many methods of experimentation on "learning" have little in common is inescapable. Thirty of the crude r's in the group of 84 summarized by Hall were corrected for attenuation due to errors of measurement, and the median r merely increased fom 0.29 to 0.47. Furthermore, the new intercorrelations reported by Hall ranged between 0.11 and 0.40 after correction for attenuation, even though all the practicable experimental precautions against spuriousness and systematic attenuation were taken.

Another instance, even more conclusive, of the opportunity for error in the definition of the function measured by a particular method is given by Commins, McNemar, and Stone (**21**), and Tomilin and Stone (**125**). In these 2 studies it was found that the function measured in the rat by the maze or problem box is almost completely independent of the function measured by the multiple discrimination box. Commins, McNemar, and Stone make a detailed and forceful plea for the experimental validation of the assumptions made regarding the functions measured by different experimental methods.

In the second place, even though both methods and materials are apparently representative of the same special class, the amount of overlap in the functions measured is frequently slight.

For example, Heron obtained (**40**) uncorrected coefficients ranging between 0.02 and 0.65 when he correlated the measurements (time, trials, or errors) obtained with 5 seemingly comparable stylus mazes that were administered 1 week apart. In the case of errors the r's were 0.23, 0.33, 0.42, 0.39, 0.35, 0.11, 0.50, 0.31, and 0.65. With nearly optimal conditions for the determination of the intercorrelation of measurements obtained when subjects learn 2 comparable stylus mazes, Spence (**110**) obtained product-moment r's of 0.60 (errors), 0.54 (trials), and 0.73 (time). Similarly, in the case of rat mazes, R. L. Thorndike (**117**) has recently reported correlations, corrected for attenuation, of 0.49, 0.86, 0.66, 0.32, 0.48, and 0.71 between the errors made in the various mazes, and Tryon (**132**) has reported the relatively high r of 0.79 (corrected for attenuation) for the errors made by rats in 2 complex and highly reliable multiple-T mazes. In the case of the common methods for studying verbal learning, the obtained intercorrelations rarely exceed 0.60, even though different forms of the "same" method and material are employed. Heron (**40**) has reported coefficients of 0.58 (time), 0.48 (trials), and 0.57 (errors) between the performances of subjects on 2 ten-letter Peterson Rational Learning Problems; and Garrison (**36**) has reported r's of 0.66 (errors), 0.45 (trials), and 0.55 (time) between a six-letter and an eight-letter rational learning problem. Finally, the relatively low correlation between different tests of immediate memory and between different tests of verbal learning is revealed by Anastasi's summary of such experiments. In short, the assumption that the *same* function is being measured when the investigator employs presumably comparable materials and methods is clearly fraught with considerable danger. In considering these intercorrelations, it must, of course, be remembered that other factors in the situation besides the material and method used may change from measurement to measurement, thus reducing the size of the coefficient obtained. Chief among these factors is the changing state of the subject (see p. 354 ff.).

The case with the different measures obtained during a single learning is no different. Although students of learning use time scores, error scores, or various combinations of these measurements in their experimental studies, it is well known that these measures are not perfectly correlated. Moreover, this lack of correlation is even more significant than in the case of the r's between measurements

obtained with different methods and materials, because the various measures are obtained simultaneously and the attenuating effect of changes in the state of the subject is eliminated.

In some cases the lack of correspondence of the different measures of learning is extreme. For example, Liggett (**65**) has reported an r of only 0.02 between the total time and total errors in the learning of a maze by rats. In human learning the obtained r's generally indicate an appreciably greater community of function, although they are still far from unity. Husband (**53**) has reported r's of 0.78 and 0.89 between trials and errors in the learning of a ten-alley U-maze by rats and human adults, respectively. In a study of the effect of visual exposure on the rate and reliability of stylus maze learning Peterson and Allison (**94**) have reported r's ranging between 0.64 and 0.89 for trials and errors, between 0.46 and 0.76 in the case of trials and time, and between 0.58 and 0.86 for errors and time. In this last instance it seems rather clear that error scores are more representative of the common factor measured by all scores than are the time or trial scores.

It is important to note at this time that low intercorrelations may have a different methodological significance in the case of measures of learning than they have in the case of the methods and materials used in obtaining the measurements. In the selection of methods and materials the choice must be between strict alternatives; whereas several measures of learning may be easily obtained during any experiment, and each measure may be used separately or as a member of a battery. Accordingly, Peterson and Allison (**94**) have raised the question whether the investigator should want the maximal correlation between the different measures of a particular learning or whether he should value more highly a low correlation between the various measures, since the group of measures would constitute a better battery for the measurement of "learning" in the latter case. A conclusion on this point need not be stated at the present time. However, the determination of the representativeness of single measures is obviously important in so far as it is desirable to limit to some extent the number of measures used. Likewise, knowledge of the most representative measure is essential for the systematic study of the intercorrelations between measurements obtained by different methods and materials (*e.g.* compare the intercorrelations obtained by Heron, **40**, and Garrison, **36**, using time, trials, and errors). This does not deny that multiple measurements may have as their chief value the more complete representation of the functions in question when they are used as a battery.

2. *Available Indices of Representativeness.* Once the need for evaluations of the extent to which particular methods for studying

learning or memory are representative of general and group abilities or factors is realized, it becomes clear that a complete methodology must evaluate different methods, materials, and measures in terms of their relative representativeness of the general or group function in question.

The studies in which Garrett (33) and Anastasi (2, 3) determined the existence of a group factor of immediate memory, and the subsequent studies by Bryan (16) and Garrett, Bryan, and Perl (35), are models of the procedure required in order to evaluate different methods, materials, and measures used in studying all types of learning and memory. Thus, Anastasi (2) discovered that a group factor was common to tests of the immediate retention of paired words, paired pictures and numbers, paired geometrical forms and numbers, paired colored forms and words, single words in series, and the recognition of nonsense syllables. She used Spearman's Tetrad Difference Method and the simpler and more reliable method of correlating each test and the common factor by the formula $r_{ag}=r_{ab}\,r_{ac}/r_{bc}$. It was then possible to determine the extent to which each test yielded measurements that were representative of the common factor, and thus to compare the tests in this respect. For example, the correlation of the paired words test with the common factor was found to be 0.66 and the correlation of the recognition test (nonsense syllables) with the common factor was found to be 0.46. The method yields, therefore, an index of the validity of any one test as a representative of a group of tests.

These methods, or those developed by Hotelling (46) and Thurstone (120, 121), provide the means of determining (a) the common factors in various types of experimental procedures, materials, and measures, and (b) the extent to which particular methods, materials, and measures represent these various common factors. They offer to the experimentalist the needed techniques for determining the validity of his methodology in precise quantitative terms. Consequently, the student of human learning may determine the number of different dimensions or factors needed for the description of all types of learning, and he may discover the particular methods, materials, and measures most suitable for the experimental investigation of these types of learning. Thus, it becomes possible to determine whether a type of learning called verbal-motor must be distinguished from other types of learning. If so, it is apparent that this type must be represented in a complete program of studies of human learning and retention. Furthermore, it becomes possible to determine whether the stylus maze, the Miles raised finger maze, the punch-board maze, or some other material is the most adequate for the study of verbal-motor learning; whether certain conditions of experimentation, types of instruction, etc., lead to more representative scores; and whether certain measures, such as trials and errors,

are more representative than other measures, such as time. A methodology built on such foundations should lead not only to an increase in the validity of the interpretations of specific experimental results, but should also aid materially in the systematization of the results of different experiments on human learning.

3. *Sources of Error in Determinations of Representativeness.* There are, however, many sources of error in studies of the community of function between measurements obtained by different methods, and these must be eliminated or reduced to a minimum if the resulting evaluation of the different methods are to be dependable.

The requirements of such studies and the important sources of spuriously high or attenuated r's in studies of learning and memory have been stated by Anastasi (**2**), Cureton (**25**), Tryon (**132**), Commins, McNemar, and Stone (**21**), and Hall (**38**). In addition, most summaries of the sources of error in the determination of reliability coefficients for learning methods (*e.g.* **64**, **110**, **130**) are appropriate, since one method of determining the reliability coefficient is almost indistinguishable from the method of determining validity within a restricted range of methods or materials (see p. 358 ff.).

In considering the significance of experimental determination of the representativeness of different methods, materials, and measures by methods that involve the correlation technique, four groups of factors must receive attention. The first of these is the statistical method employed. There are a number of methods that may be used in determining the existence of common factors and they are not equally serviceable or precise. The crudest method is, of course, merely the determination of the average correlation between measurements obtained by a single experimental method and the measurements obtained by all the other experimental methods used. In this case the r's should be converted into Fisher's z-function before averaging (**30**, p. 188).

Although this method is the one most commonly used in studies of learning methods, it is at best only suggestive (**2**). Nevertheless, it has received empirical validation in studies of the immediate memory factor conducted by Anastasi (**2**, **3**), Bryan (**16**), and Garrett, Bryan, and Perl (**35**). That is, the factors indicated by the average intercorrelations have also been indicated by the Tetrad Difference Method, by the Thurstone Method of Multiple Factor Analysis, and by the method that involves the determination of the correlation between each type of measurement and the common factor (r_{ag}). As for the more complicated methods, there seems to be no difference in the precision of the results obtained (**34**). However, Anastasi (**2**) has pointed out that the Tetrad Difference Method may distort the results if the correction of the crude r's for attenuation is large, and that the tetrad criterion may be satisfied even though the correlation between one of the tests and the common factor is zero.

Furthermore, the Thurstone Method of Multiple Factor Analysis has greater generality than the Method of Tetrad Differences (120), since the latter is merely a special case of the former. The Method of Multiple Factor Analysis is probably the most adequate for determining the minimal number of independent factors that must be postulated in order to account for the correlations between the measurements obtained by a number of *unselected* methods.

The second factor that must receive consideration is the sample of subjects used. (a) The intercorrelations are most interpretable when the subjects represent accurately a range of talent in a homogeneous population. The obtained *r*'s are attenuated when the range of talent is restricted, because a correlation coefficient represents the ratio between the variance attributable to true individual differences and the total variance, *i.e.* the variance determined by the true individual differences plus errors of measurement (129). Many of the low correlation coefficients reported in the studies of memory and learning methods are undoubtedly attributable to such a restriction of the range of talent sampled. On the other hand, if a heterogeneous population is sampled, *i.e.* the individuals differ in important traits such as sex, age, educational status, etc., the *r*'s may be either raised or lowered, depending upon whether the 2 abilities correlated are positively or negatively correlated with the irrelevant factor.

For example, since age is positively correlated with learning ability, at least until age 25, a marked difference in the ages of the subjects used should produce spuriously high intercorrelations of the measurements made with 2 or more learning methods, materials, or measures. An important demonstration of the effect of such heterogeneity on the size of the correlation coefficient, and a demonstration of the statistical analyses necessary for the segregation of such effects, has been given by Tryon (131) in a study of the proportions of the total variance in maze scores attributable to sex, age, and diet differences among the white rats used as subjects. In general, it may be said that the failure to obtain an entirely homogeneous sample, or the failure to sample the entire range of ability within this homogeneous population, does not invalidate comparisons of the intercorrelations obtained from the same group of subjects. However, the failure to use a well-defined, representative, and homogeneous group of subjects interferes with inter-experiment comparisons of the obtained *r*'s (see p. 363).

(b) Garrett, Bryan, and Perl (35) have recently reported that the average intercorrelation of 6 memory tests and 4 non-memory tests decreases progressively from age 9 to age 12 and from age 12 to age 15, thus indicating a greater specificity of abilities as age increases. In view of this, it is apparently essential that studies involving the determination of common factors be made for several representative levels of maturity in the human subject, and that

generalizations regarding the existence of common factors and the representativeness of particular measurements be made specific to the age level of the subjects employed.

The third major consideration in the interpretation of the correlation coefficients obtained in studies of the community of function between different methods is the reliability of the individual measurements. The reliability coefficient obtained for any method, material, or measure fixes the limit of the intercorrelations that may be obtained with that method, material, or measure. That is, the occurrence of errors of measurement brings about an attenuation of the "validity" coefficient. This circumstance is particularly important in experimental studies that are designed to select the most representative experimental methods, because the true representativeness of a particular method may be very high and yet appear to be low because the reliability of the measurements obtained with that method is low. Since the reliability of the measurements may be increased without altering the essential form of the method, *e.g.* by lengthening the task or by improving the environmental controls, the "validity" measurements should be freed of the effects of errors of measurement.

The usual procedure is to correct for such attenuation by the Spearman formula $r_{\text{true ab}}=r_{ab}/\sqrt{r_{aa'}}\sqrt{r_{bb'}}$. This formula is widely used among students of learning methods, and still more widely recommended. However, it does not always give an entirely satisfactory solution to the problem in the case of learning methods. There are several commonly used methods by which the reliability coefficient ($r_{aa'}$, $r_{bb'}$) may be computed, and the methods yield characteristically different coefficients (**110**); none of the methods is entirely above suspicion from a theoretical and experimental standpoint (see p. 364 ff.). Yet the amount of the correction for attenuation obviously varies inversely as the magnitude of the reliability coefficients employed.[2] Thus, if the reliability coefficients are computed by using the summations of the errors made on odd and even trials, high r's are usually obtained and the correction of the "validity" coefficient is small. On the other hand, if the reliability coefficients are determined by correlating scores on learning and relearning, low r's are obtained, and the correction of the "validity" coefficient is large. This problem can be solved, and then only partially, in no other way than by increasing the reliability of measurements as much as possible before attempting to determine the intercorrelations of measurements obtained with different methods.

The fourth consideration in evaluating the intercorrelations of measurements obtained with different methods concerns the extent

[2] Cureton (**25**) has presented formulae for checking the assumption that the reliability coefficients involve no correlation of errors of measurement. The formulae used by Brown (**12, 13**) are considered inadequate.

to which the subjects have been "held constant" during the experiment.

> Any variation in the state of the subject, or any modification of the behavior of the subject during the performance of one task that conditions his performance in the next task, leads to a deviation of the obtained r from the true r, or rather, from the r that should be obtained if chance factors were the only attenuating causes. In learning experiments, the major sources of inconstancy are the basic changes in the ability of the subjects from day to day (quotidian variability), changes in motivation from task to task, changes in the subjects' habituation to the laboratory environment, and specific positive or negative transfer from task to task. These factors may be considered most effectively in the discussion of the sources of error in the determination of reliability coefficients (pp. 353–364).

We have considered at length a statistical method of evaluating different experimental methods in terms of one criterion of validity, namely, the extent to which a method, material, or measure yields measurements that represent statistically defined factors in learning. This statistical method has the advantage that it provides objectively defined types of learning or learning factors and a quantitative index of the extent to which measures obtained with different methods represent these types of learning. On the other hand, the disadvantages of the statistical approach are easily discerned. As an evaluative technique of importance in developing a methodology of human learning, the statistical method rests on the assumption that the true community of function between different series of measurements may be estimated when the obtained coefficients are known. This assumption holds only when the obtained coefficient deviates from the true coefficient as a consequence of sampling errors, or of uncorrelated chance errors of measurement. But it is clear that this assumption is frequently unwarranted, and in studies of learning methods the sources of spuriously high or low correlations may not only fail to yield to statistical corrections, but may also defy experimental control. There is, therefore, the danger that many misleading analyses of the factors in learning and many incorrect evaluations of the methods of studying learning may result from the use of the correlation coefficient. Finally, all the aspects of a complete methodology cannot be evaluated with reference to their validity in this way.

4. *Direct Experimental Analyses of Representativeness.* The alternative method for determining the representativeness of any aspect of an experimental method used in the study of learning is the traditional method of direct experimental analysis. This method needs no formal statement. It is obvious that the extent to which a

particular measure represents learning may be determined by a critical examination of the activities of the subject that lead to a change in that measure. Likewise, the characteristics of a learning material may be ascertained by an experimental analysis of the reactions of the subjects to that material, and the extent to which certain methods of control permit the measurement of learning in isolation from other confusing and irrelevant activities may be determined by an analysis of the behavior of the subjects when that procedure is used.

For example, the extent to which nonsense syllables (*VOJ*) and three-consonant units (*VBJ*) are valid representatives of meaningless verbal material may be determined experimentally. In fact, Glaze (**37**), Hull (**47**), and Krueger (**61**) have determined the number of associations aroused by nonsense syllables, and Witmer (**141**), after quantifying the meaninglessness of three-consonant units in the same way, has shown that the three-consonant units are more representative of meaningless material than nonsense syllables. Likewise, it has been found by Warden (**138**), Husband (**54**), and others that stylus and finger mazes are learned by most human subjects by methods that involve verbal self-direction and verbal trial and error, and that the maze cannot be considered as a device for the study of "pure" motor habits in the human subject. Such experimental analyses of the validity of various methods, materials, and measures have been made since the beginning of the experimental study of learning. Moreover, every experimental analysis of the factors that condition learning is a possible source of valuable data pertaining to the validity of the methods of experimentation employed.

B. Freedom from Sources of Experimental Error as a Criterion of the Validity of an Experimental Method

Most criticisms of experimental studies emphasize uncontrolled sources of experimental error rather than the inadequacy or inaccuracy of the investigator's identification of the basic factors operating throughout his experiment, *i.e.* identification of the type of learning represented in the experiment. To return to the example cited earlier, the investigation of the relationship between the amount of muscular tension and the speed with which a stylus maze is learned is subject to the most severe criticism if it permits a constant difference, other than the amount of muscular tension, between the conditions under which the 2 groups of subjects learn. On the other hand, if the investigator concludes that he has discovered the influence of muscular tension on *motor* learning, and no constant experimental errors are apparent, critics may point to the need for a correction or restriction of this generalization so that it reads *verbal-motor learning* or *stylus maze learning*, but the experimental results

will be incorporated into the body of accepted knowledge. In short, the identification of the class of operations to which the particular experimental operations belong increases the importance of the isolated experimental results but does not guarantee the absence of constant experimental errors, nor is such identification necessary for the acceptance of the isolated experimental results as valid. Therefore, the investigator not only needs methods, materials, and measures that are representative of a well-defined class of methods, materials, and measures, but he also needs experimental methods, materials, and measures that are least likely to permit unrecognized constant differences between experimental conditions. The presence or absence of these unrecognized constant differences determines the truth or falsity of any generalization, whatever its scope, regarding the relationship between experimental variable X and experimental result Y.

The process of inductive generalization from the data of an experiment is one that yields to no formal all-or-none criterion of truth or falsity. It is clear that the statistical methods for determining the reliability of an obtained difference cannot serve as criteria of the presence or absence of unidentified constant differences between experimental conditions.

As Boring (**10, 11**) has so clearly indicated, the application of the statistical test of the reliability of an obtained experimental difference is in the nature of a more exact description of the obtained data. The statistical tests of significance merely answer the question: Is it possible to explain this difference as the result of chance? If the difference is less than three times the $\sigma_{diff.}$, and the data have been obtained under conditions that satisfy the assumptions made in estimating the $\sigma_{diff.}$ (see p. 337 ff.), the investigator knows that the difference *may* have been the result of a particular combination of variable errors, and he may obtain further data in an effort to test this hypothesis. But these statistics do not and cannot reveal, regardless of the number of observations made, whether the obtained difference is too small as a representation of the effect of experimental variable X, or whether the obtained difference is too large as a representation of the effect of X. Such generalizations regarding the effect of X are conditioned by the amount of the obtained experimental difference and by its apparent reliability, but the final induction rests on the judgment of the investigator regarding the success of his attempts to equilibrate all the basic experimental conditions before introducing variable X. The lack of dependence of this judgment on the consideration of mathematical relationships does not signify that it is an expression of preconceived ideas as to what should happen when variable X is introduced; it is determined by the degree of confidence of the investigator in the experimental methods and controls used in the investigation. In the final analysis, judgment of the validity of an experimental difference as an indicator of the effect of an experimental variable is

based on the apparent number of possible sources of error in the experiment, or on the apparent failure to control certain factors that other studies have shown to be important.

The probability of constant experimental errors and of invalid generalizations is determined, therefore, by the experimental methods, materials, and measures used, and *the relative validity of different methods, materials, and measures may be determined in terms of criteria that refer to the possibility of obtaining inequivalent basic conditions of experimentation when they are used.* Formal quantitative criteria of this type are elusive. However, three simple criteria may be employed in evaluating the various aspects of a methodology.

(1) *The number of contradictory, yet statistically reliable, conclusions obtained with a particular method may be used to judge the extent to which unrecognized constant variations in the basic conditions may occur when that method is used.* Thus, if contradictory conclusions regarding the effect of a certain variable on the speed of memorization are obtained when the method of complete presentation is used under seemingly comparable conditions by different experimenters, and no contradictory conclusions are obtained when the effect of the same variable is studied by means of the anticipation method, it is legitimate to conclude that the anticipation method provides a more adequate control of potential sources of error than the method of complete presentation. This criterion is obviously indistinguishable from that of reliability unless the contradictory conclusions have been supported by statistically reliable differences. Few are the opportunities to apply this criterion in evaluating methods used in the study of learning, because so few experiments have been repeated. Nevertheless, this criterion is valuable as a formalization of that which is implied when investigators refer to "experience" in judging a method of investigation.

(2) If it be granted that the probability of an incorrect interpretation of an experimental difference increases as some function of the number of uncontrolled aspects of the experimental situation, it may be concluded that *an experimental method is more adequate the greater the positive control over all factors in the experimental situation and the more complete the measurement of those factors that cannot be held at absolutely constant values.* Thus, constancy of the duration of exposure of memory materials is considered superior to a method in which the time of exposure is determined by the idiosyncrasies of each subject, since the control of the duration eliminates one possible source of a constant unintended difference between the

conditions of an experiment. The same example may be used to illustrate the second part of the proposed criterion. The use of a method that permits each subject to determine the rate of exposure is more acceptable if the experimenter measures the rates of presentation used by each subject, because it is then possible to determine whether a constant difference between the experimental conditions with respect to this factor actually occurs. There is, of course, a limit to the application of this ideal of experimental control, if an experiment is ever to be performed. However, it will be found useful in the evaluation of different experimental methods in common use, since the principle has been frequently neglected.

(3) Whether any particular aspect of the experimental situation *needs* positive control may be determined by specific methodological studies such as the one conducted by McGeoch (**82**) on the effect of the experimenter on the performance of subjects in the learning and relearning of adjectives. Furthermore, the relative validity of the assumptions involved in the use of different methods of experimental control or different learning materials may be determined by experimental analysis. Thus, the validity of the various methods used to equilibrate practice effects in experiments in which the subjects are used under more than one condition may be determined by studies of the accuracy with which constant differences in practice are eliminated when no experimental variations in the conditions are introduced.

II. THE CRITERION OF RELIABILITY

The second group of criteria that may be used to judge the relative adequacy of different experimental methods for the study of human learning refers to the reliability, or dependability, of the experimental measurements obtained when those methods are used. One of these criteria, namely, that *the experimental method is more adequate the smaller the variable error in the measurements obtained when it is used,* received its first formal statement in terms of an index of variable error when Spearman (**108, 109**) defined the reliability coefficient. Although some of the early applications of this index in the evaluation of mental tests included the application of the criterion to certain tests of learning (*e.g.* **12**), Hunter (**49**) was the first to apply the criterion in the evaluation of the materials and methods that were being used in experimental studies of the general phenomena of learning. The past 15 years have witnessed a marked increase in the frequency with which this criterion has been applied

in evaluating the methods used to study learning, particularly by students of animal learning, and an increase in the number of different ways of determining the amount of variable error involved in measurements. Thus, the reliability coefficient has been supplemented in recent studies by direct measurements of variability as represented in the $\sigma_{dist.}$, and by direct determination of the amount of variation between the means of measurements obtained under the same experimental condition. However, complete validation of the indices of variable error proposed, and standardization of the experimental situations used in comparing different experimental methods in terms of these indices have not been achieved.

The emphasis of Hunter and others on the measurement of the variable error attributable to the use of particular methods has carried with it the neglect of another important criterion in terms of which an experimental method may be evaluated, namely, that *a method is more adequate the greater the possibility of an accurate statistical analysis of the variable errors in the data obtained when that method is used.* This criterion has not yet been employed by students of the methodology of learning, but its usefulness is suggested by the fact that it represents a general criterion of the adequacy of experimental methods. The applicability of this criterion to the evaluation of psychophysical methods is clearly recognized by Culler (**24**), and a formal statement of the general principle has recently been given by Fisher (**31**) in his treatise on *The Design of Experiments.* Before considering in detail the two criteria mentioned above, it will be profitable to relate them to the usual logic employed by investigators in arriving at an estimate of the dependability of their specific experimental results.

The two criteria that have been mentioned will be recognized as axiomatic when the process whereby the investigator determines the dependability of his isolated experimental results is examined. The customary procedure in an experiment is to make a number of observations under Conditions X and Y, summarize the distributions of these observations in terms of measures of central tendency and variability, and then judge the dependability of the obtained (say) mean difference by comparing the mean difference between the measurements for Conditions X and Y with such differences as might be expected between these means in view of the observed variations in the measurements obtained under like conditions (*i.e.* σ_X and σ_Y). The accepted formula for making this comparison is:

$$\text{C.R.} = \frac{\text{Mean Difference}}{\sqrt{\left(\frac{\sigma_{\text{dist.X}}}{\sqrt{N}}\right)^2 + \left(\frac{\sigma_{\text{dist.Y}}}{\sqrt{N}}\right)^2 - 2r_{XY}\left(\frac{\sigma_{\text{dist.X}}}{\sqrt{N}}\right)\left(\frac{\sigma_{\text{dist.Y}}}{\sqrt{N}}\right)}}$$

The critical ratio, or the ratio of the mean difference to the sigma of the mean difference, is then interpreted in terms of the probability that the obtained difference between the X and Y observations may have occurred as a consequence of the errors of measurement, or variable errors, similar to those observed in the X and Y values. The application of this statistical test of significance implies that the investigator has the hypothesis that the X and Y values belong to the same normal universe, and his hypothesis is usually considered to be disproved when the obtained mean difference is 3 times its own sigma. If the critical ratio is less than 3:1 the investigator who accepts this standard of significance must conclude that he has failed to disprove his original hypothesis. It is of considerable importance to recognize the fact that he cannot under such circumstances draw the conclusion that the measurements obtained under Conditions X and Y are from the same population, *i.e.* that Conditions X and Y are not essentially different. As Fisher so clearly states, "The null hypothesis is never proved or established but is possibly disproved in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis"(**31**, p. 19).

The Reduction of Variable Errors as an Aim of Methodological Research. An analysis of this paradigm of experimental method reveals several important facts regarding the relationship between the experimental method and the opportunity presented to the investigator to make a valid and productive statistical test of his hypothesis. It is apparent that the reliability of the experimental method is of importance in determining the sensitivity of the experiment as a test of the investigator's hypothesis. In view of the relationship expressed in the formula $\sigma_{\text{mean}} = \frac{\sigma_{\text{dist.}}}{\sqrt{N}}$, the experiment may be made more sensitive, *i.e.* capable of detecting a smaller departure from the "null" hypothesis, either by increasing the number of independent observations under Conditions X and Y or by reducing the errors of observation or variable errors present under Conditions X and Y. Since the variability of measurements under Condition X (and likewise Condition Y) indicates that there has been a variation from measurement to measurement in the underlying condition that we

call X, it follows that the variability of measurements obtained under Condition X may be reduced by refining the techniques of experimentation used. We are led, therefore, to the statement of the first criterion for evaluating an experimental method, material, or measure: *The method, material, or measure is more adequate the less the amount of variable error involved in the measurements obtained when it is used.* The reliability of an experimental method is an inverse function of the associated variable error.

It is appropriate at this time to consider the question of the relationship between the reliability of a method, material, or measure and the proper formula to be used to provide an estimate of the validity of the hypothesis that the obtained experimental difference is a chance difference. In the form in which we have stated the criterion that refers to the variable errors involved in the use of a method, it is assumed that the chief value of a reliable method is the economy of experimental labor—the greater the reliability of the method the fewer the number of observations required in order to make an experiment sufficiently sensitive to provide a check on the investigator's hypothesis. However, when Hunter (**49**) first proposed the use of the reliability coefficient as an index of the variable errors involved in the use of the maze in learning experiments, he argued that an unreliable maze ($r<+0.30$) was not only unsuited for use in studies of individual differences, but that it was likewise unfit for studies of group differences. The implication was that mazes were so unreliable that group differences were not dependable even though they satisfied the usual statistical test of significance. Thus, he said, "Scientific dependability is not secured by dumping a number of unreliable measurements together in one group" (**52**, p. 441). Carr (**18**) rightly objected to this argument for more reliable methods on the ground that it denied the validity of the very thing that experimentalists are forced to do at all times, and pointed out that the usual statistical test of significance evaluates the obtained differences in terms of the variable error involved in measurements under like conditions. In short, Carr affirmed the value of developing more reliable methods, but denied that such improvements were *necessary*. Carr's defense of the orthodox statistical test of the significance of a difference was later accepted by Hunter (**50**) and elaborated by Tolman and Nyswander (**123**).

Following this apparent solution of the problem regarding the relationship between the reliability of a method and the validity of the statistical test of significance of experimental differences, Tryon (**126, 127**) contributed a statistical proof that Hunter's original contention was incorrect, but in the course of this proof was led to conclude that the usual formula underestimated the true dependability of an obtained difference in proportion as errors of measurement occurred. Tryon proposed the following formula as the "proper" one to use in testing the significance of an obtained experimental difference:

$$\text{Critical Ratio}_{\text{true}} = \frac{\text{Mean Difference}}{\sqrt{\left(\frac{\sigma_X\sqrt{r_{XX}}}{\sqrt{N_X}}\right)^2+\left(\frac{\sigma_Y\sqrt{r_{YY}}}{\sqrt{N_Y}}\right)^2}}$$

The chief difference between this formula and the orthodox formula is that "true" σ's ($\sigma_{true}=\sigma_{dist.}\sqrt{r}$, where r=reliability coefficient) are used instead of the fallible obtained σ's. At this time he concluded, "It is readily conceivable that in many actual experiments, differences between groups have been obtained which, by the ordinary use of I_σ [the usual critical ratio formula], indicated that these differences could easily have occurred by chance, whereas a correction for the unreliability of measurement would have shown that these differences could most improbably have arisen by chance" (**127**, p. 22). This conclusion does not deny the importance of improving the reliability of methods, since an underestimation of the reliability of experimental results is clearly as undesirable as an overestimation of their reliability. Nevertheless, the formula apparently rests on assumptions that are not sound. In a later paper Tryon (**128**) fails to retract the original recommendation to investigators but concludes with the statement that "for any given two samples of subjects undergoing measurement whose true sigma difference due to errors of sampling may be therefore considered a constant with reference to other types of errors, the greater the sigma difference due to errors of individual measurement, the greater will be the total fallible sigma difference computed by the orthodox sigma difference formula, *and hence the less significant the difference found*" (**128**, p. 195). It is probable that Leeper (**64**) has indicated correctly the source of error in Tryon's argument, namely, that Tryon assumed that the difference between means obtained from different samples of subjects would be relatively unaffected by the magnitude of the errors of measurement, since he made no correction of the numerator in the usual critical ratio formula. This assumption is unwarranted. The essence of the statistical argument behind the use of the critical ratio is that as the variability of measurements under Conditions X and Y increases, the possibility of obtaining large differences between the means of X and Y by chance alone likewise increases.

It must be concluded that the usual formula for the statistical test of significance is valid in so far as it compares accurately the mean difference between measurements obtained under different conditions with such differences as might be expected between means in view of the observed variations (from whatever causes, sampling errors or errors of measurement) obtained under like conditions. The use of Tryon's formula leads to an *overestimation* of the reliability of obtained differences because it removes something from the statistical estimate of error without actually removing it as a source of error in the experiment. The chief value of a reliable method is the economy of experimental labor. The permissive character of this consideration does not lessen the actual importance of reliability as a criterion for the selection of experimental techniques.

There are practical limitations to the number of observations that can be made in most experimental studies. This limit is particularly low in the field of learning and memory, since the single observations are often lengthy and are more often enmeshed in a complicated

schedule designed to counterbalance changes in the subject such as practice and fatigue effects. As recently pointed out by Corey (22), there are few experiments on learning and memory in which more than 25 observations have been made under each experimental condition, and the number is far less than this in the highly significant comprehensive experiments in which several factors have been systematically varied. In addition, it is frequently impossible to obtain repetitions of measurements that are independent of those already made, and the formulae that are used to estimate the error in the mean from the known $\sigma_{dist.}$ and N assume that the observations are independent. It is for this reason that the investigation of individual differences is frequently said to require reliable techniques, since the repetition of the test must use the same individuals, and the changes produced in the organization of the individual during the first test are carried over and determine to some extent the reactions to the second test.

Improvement of the Accuracy of Statistical Estimates of Error as an Aim of Methodological Research. Further consideration of the assumption that the measurements included in a single series are uncorrelated leads to the second criterion in terms of which experimental methods may be evaluated. This second criterion refers not to the actual extent to which variable errors have been eliminated but to the extent to which the variable errors may be accurately estimated in making a statistical test of the significance of an obtained experimental difference. When the investigator has not performed the experiment in such a way that an *accurate* estimate of the variable errors may be made in the course of applying the statistical test to his " null " hypothesis, the application of the test becomes of doubtful value. Thus, if the usual test gives an estimate of error that is appreciably too high or too low, the investigator may be led falsely to conclude that he has either disproved or failed to disprove his hypothesis that the obtained difference is a chance difference. Either eventuality reduces the validity and precision of the conclusion, since it nullifies the significance of this preliminary stage in the process of induction. Nevertheless, many experimental methods in common use have within their structure the sources of unrecognized nonrandom error and unrecognized restrictions of error that are unmeasured and are consequently not taken into account in the statistical test of the experimental finding. Various instances of this characteristic of certain experimental methods have been recognized recently by Culler (24), Woodrow (143), and Fisher (31, p. 88 ff.).

The generalized form of the criterion that results from such considerations, as previously stated, is that *a method is more adequate the greater the possibility of an accurate statistical analysis of the variable errors in the data obtained when that method is used.*

The full implication of this criterion is that, as Fisher has said (**31**, p. 89): "The consequences of accepting an insignificant effect as significant, or of rejecting as insignificant one which, with sounder methods of experimentation, would have shown itself to be significant, are equally unfortunate. In fact, the calculation of standard errors is idle and misleading, if the method . . . adopted fails to guarantee their validity, and the same applies to all other means of testing significance." Yet it may be fairly stated that many of the methods used in the study of learning fail to insure an *accurate* estimate of the variable error for use in interpreting the results, although investigators usually attempt to make certain that the estimate of error is not an *underestimate*. For example, it is not uncommon for investigators to use the same subjects or matched groups of subjects in the various conditions of an experiment but fail to correct for this restriction of the error of sampling by using the proper formula for the standard error of the difference. The attitude appears to be that the estimate of error obtained without taking into account the fact that certain sources of error have been restricted is not only acceptable but even desirable, because the investigator will then be certain that it is not an underestimate and that the dependability of the obtained difference is not overestimated.

This practice has been frequently criticized (**66**, **92**, **135**). From the standpoint of the logic of experimentation it involves a perverted emphasis on the disproof of the hypothesis that chance may account for the obtained experimental differences, and assumes that science is concerned only with the question whether factor X has an effect; whereas, if the systematization and quantification of the results of experiments is a legitimate aim, the statistical and experimental methods must be adequate for determining not only whether factor X has an effect, but also whether factor X has an effect of a certain magnitude (**31**, p. 190). Thus, the fact that factor X produces a mean difference of 10 trials in learning, and that the standard error of this mean difference is 1, has a significance beyond that of indicating that factor X produces a change in measurements which is dependably greater than zero. It also indicates that factor X produces a change of at least 7 trials but not more than 13 trials in learning. Unless the methods used to estimate the chance errors involved in the experiment yield neither underestimates nor over-

estimates of the $\sigma_{\text{mean diff.}}$ this important refinement of the interpretations of experimental results cannot be attempted. From the standpoint of statistics the practice of ignoring the extent to which the variable error in the experimental measurements has been restricted involves the fallacy of using what Fisher (30, p. 12) has called *inconsistent statistics, i.e.* ". . . a statistic which even from an infinite sample does not give the correct value; it tends indeed to a fixed value, but to a value which is erroneous from the point of view with which it is used."

Important advances have recently been made in the development of statistical formulae that permit an accurate estimate of error in instances in which particular experimental methods have been employed. Thus, Lindquist (66, 67) and Peters and Van Voorhis (92) have developed formulae for use in experiments in which the subjects have been matched with respect to some criterion factor. In effect, these formulae permit the investigator to determine his statistical estimate of error after having taken into account the reduction in the sources of error that has accompanied the use of a particular experimental method. These writers voice justified criticisms of the investigator's use of inconsistent statistics and direct their efforts toward providing consistent statistics for use with certain experimental methods and toward making the investigator aware of the fact that he must use a statistical test that is appropriate to his experimental method.

The emphasis may, however, be reversed. As will be shown later, there are some experimental methods which yield data that cannot be appropriately analyzed with any of the usual statistical tests, and it is doubtful whether appropriate corrections for the usual statistical formulae will be forthcoming. Therefore, it is appropriate to require that an experimental method for the study of learning not only yield minimal variable errors, but also yield data that are susceptible to accurate analysis by the available statistical methods for estimating the range of chance error.

In the succeeding sections the validity of these two criteria is assumed, and the discussion centers around the problems involved in applying these criteria in the actual comparison of experimental methods used in the study of human learning. A particular problem in each case is that of determining valid quantitative indices in terms of which the comparisons may be made.

A. The Availability of Accurate Statistical Estimates of Error as a Criterion for Evaluating Experimental Methods

The investigator employs the usual formulae for the σ_{mean}, $\sigma_{\text{mean diff.}}$, and other estimates of error on the assumption that the measurements included in his sample have been collected under con-

ditions that Yule (**146**, p. 259 ff.) has termed "simple sampling," and that Fisher (**31**, p. 20 ff.) has termed "randomisation." If the experimental method has been such that these conditions are satisfied, the σ_{mean} computed by the formula $\sigma_{dist.}/\sqrt{N}$ is an accurate estimate of the standard deviation of a distribution of means from like samples with the same N. Likewise, the $\sigma_{mean\ diff.}$ computed by the usual formula is an accurate estimate of the standard deviation of a distribution of mean differences between Conditions X and Y such as would be obtained if the experiment were repeated a number of times without change. On the other hand, when the conditions of simple sampling have not been satisfied, the estimated σ_{mean} and $\sigma_{mean\ diff.}$ are greater or less than the standard deviations of the distributions of means or differences between means that would be obtained if the experiment were repeated. It is clear that the σ's of the actual distributions of means and mean differences are the "true" measures of the reliability or dependability of obtained means and mean differences, and that the statistical formulae provide, under appropriate conditions, merely an estimate of what would be obtained if investigators were in the habit of repeating their experiments. Therefore, the proper index of the accuracy of the statistical estimates obtained when the investigator uses a particular experimental method must be some index that expresses the relationship between the σ_{mean} predicted on the assumption of "simple sampling" and the actual distribution of means obtained when the observations are repeated.

1. *Indices of the Accuracy of Statistical Estimates of Error.* Woodrow (**143**) has recently proposed 2 such indices in a study of the quotidian variability of human subjects. His problem was to determine whether the differences between the means of groups of measurements obtained from the same subject on different days were greater than the differences to be expected if chance factors, such as produced the variations in measurements during a single day, were the sole determinants of the differences.

Woodrow used as one measure of this divergence from chance the ratio between the average difference between the means obtained on different days and the average difference to be expected if chance factors alone were operating, as computed by the formula, $2\sigma/\sqrt{\pi N}$ or $1.1284\sigma/\sqrt{N}$, in which N is the number of measurements comprising one day's sample and σ is the standard deviation of the total population of measurements (all days combined). The ratio of these 2 values is unity when the measurements on any one day are unbiased, or random, samples from the total population of measurements, and rises above unity (5.7 in one instance) when the measurements obtained on any one day represent only a restricted portion of the total distribution, *i.e.* when

factors other than chance are operating to determine the measurements obtained on different days. This same index could, of course, be used to detect an undernormality of the distribution of means. However, Woodrow favors another index that gives practically the same ratio values when used to analyze the same data, because this second index employs the more customary statistic of dispersion, σ, and is less laboriously calculated. This index, named the *index of quotidian variability*, is the ratio of the standard deviation of the means obtained on different days and the average of the σ_{mean}'s estimated by the usual formula, $\sigma_{mean}=\sigma_{dist.}/\sqrt{N}$, from the known variability of the measurements obtained during each day.

The index of quotidian variability is one form of a more general index used by Lexis in 1877 (**136**, p. 45; **98**, p. 87 ff.) to investigate the general question of the extent to which particular sets of observations conform to the law of error. The first application of the Lexian ratio in the study of mental phenomena was made by Culler (**24**) in an empirical determination of the appropriate formula for the *P.E.* of the constant process limen in the case of lifted weights. The only important difference between the index of quotidian variability and the Lexian ratio is the use of the average of the estimated σ_{mean}'s from the single samples in the denominator of the ratio in the former case and the use of the single estimated σ_{mean} in the latter case. Although the use of the average of the estimated σ_{mean}'s is to be preferred when the index is employed in methodological investigations, the index will be referred to as the Lexian ratio since Woodrow's term implies a restriction of the usefulness of the ratio to studies of overnormal distributions of means from successive samples, whereas an equally important application of the ratio is in the discovery of experimental methods that yield undernormal distributions of means from successive samples, *i.e.* methods that permit the underestimation of the reliability of obtained means and mean differences.

The Lexian ratio is obviously the quantitative measure needed for the evaluation and comparison of different experimental methods in terms of the applicability of the usual statistical methods for estimating error.

Whenever L differs from unity by a significant amount the statistical analysis based on the assumption of random sampling is inappropriate.[3] If L is greater than 1 the estimated reliability of the mean or difference between means is an overestimate; if L is less than 1 the estimated reliability of the mean or differ-

---

[3] The Probable Error of L, as given by Culler (**24**) and Rietz (**98**), is $.4769\ L/\sqrt{n}$, where n gives the number of samples or sets of values. Woodrow (**143**) gives the following formula for the P.E. of the index of quotidian variability:

$$\text{P.E.}_{B/A}=.6745\sqrt{\left(\frac{Ba}{A}\right)^2+\frac{b^2}{A}}$$

in which B and A are the numerator and denominator of the ratio, and b and a are the standard deviations of B and A.

ence between means is an underestimate. In either case the ratio indicates the need for (a) a change in the statistical formulae used for the analysis of error in case such more appropriate formulae are known; (b) the empirical determination of the proper correction for the usual formulae; or (c) a change in the conditions or methods of experimentation in order to make them conform more closely to the conditions of random sampling. For example, instead of the simple formula for the $\sigma_{\text{mean diff.}}$ that assumes complete randomisation of variable errors, including errors of sampling, the investigator may need to use a formula for the $\sigma_{\text{mean diff.}}$ that takes into account the fact that the subjects used under Conditions X and Y have been matched. Or, if no adequate statistical formula is known, the investigator may determine empirically that the proper formula for the σ_{mean} under his experimental conditions is, say, $\sigma_{\text{mean}}=\sigma_{\text{dist.}}/\sqrt{1.75N}$, rather than the usual $\sigma_{\text{dist.}}/\sqrt{N}$. Culler (**24**) developed such a special formula for the *P.E.* of the constant process limen for lifted weights. However, this method of overcoming the inaccuracy of the usual estimates of error has only a limited applicability in the field of learning and memory. It requires the fractionation of a relatively large number of observations in order that reliable Lexian ratios may be determined, and such large numbers of observations can rarely be obtained in studies of learning and memory other than immediate memory. Therefore, the most promising way out of the difficulty indicated by high or low Lexian ratios seems to be the alteration of the experimental methods used and the selection of those methods that yield Lexian ratios that are not significantly different from unity.

2. The Relation Between the Lexian Ratio and the Conditions of Experimentation. Although the Lexian ratio must remain the final test of the accuracy of statistical estimates of error obtained with particular experimental methods, the consequences of deviations from the random sampling procedure may be predicted with considerable accuracy. Therefore, the evaluation of many experimental methods need not rest on actual determinations of Lexian ratios.

The conditions of sampling, or experimentation, that yield Lexian ratios greater and less than unity have been adequately described by Culler (**24**, p. 467) in terms of urn schemata: "To get the meaning of L let us resort to the usual urn-schema. Given 10 urns of white and black balls in differing proportions; a set (s) of 10 is drawn and the percentage of whites (p_w) recorded; n sets (a total of ns draws) are completed. Three typical procedures are open. (i) We may always draw from the same urn (identical composition throughout the series of ns trials); the n values of p_w thus drawn will, apart from casual irregularities, constitute a binomial or Bernouillian series $(p+q)^s$, whose Lexian ratio is unity and dispersion 'normal.' (ii) We draw a set of 10 from urn 1, another set from urn 2, and so on (constant proportion within a given urn, but changing from set to set); the n values of p_w now form a Lexian series, wherein $L>1$ and dispersion 'over-normal.' (iii) We make up a set of 10 by drawing 1 ball from each urn (unlike probabilities of white from draw to draw, but the same combination of probabilities within each set); these

n values of p_w compose a Poisson distribution, whose variance is lowest of all; $L<1$ and dispersion 'under-normal'."

It is apparent that an accurate estimate of the error in the mean of a single sample is obtained only when the observations included in the sample are taken at random from a homogeneous population. When the Lexian ratio is greater than 1 the σ_{mean} computed on the assumption of random sampling may be considered as an accurate estimate of the error in the determination of the mean for the particular population sampled (urn 1, urn 2, or urn 3, etc.), but it does not measure the probable deviation of the obtained mean from the true mean of all the urns in the series. The implication of this for the investigator is too well known to need elaboration. Thus, the standard error of the mean trials required by a group of subjects to learn a first list of nonsense syllables is an accurate estimate of the probable deviation of the mean from the true performance of the subjects at that stage of practice, but it has long been recognized that it is not a measure of the probable deviation of that mean from the mean performance of the subjects on, say, 10 lists of nonsense syllables that are learned in succession. In short, the Lexian ratio greater than 1 merely represents a failure on the part of the investigator to hold the basic conditions of experimentation, including the state of the subject, constant from one set of observations to the next, and the error involved in the use of the usual statistical estimates of reliability when L is greater than 1 is chiefly a matter of the failure to recognize these constant differences in the conditions of experimentation that we have previously termed *constant experimental errors* (p. 329 ff.). The methodological significance of Lexian ratios greater than 1 is that they point to the need for such alterations of the experimental method as may be necessary to provide for the elimination or equalization of such errors under the several conditions of the experiment.

The essential characteristics of the experimental method that leads to Lexian ratios less than 1 are (a) the inclusion in the sample of measurements from essentially different populations or essentially different parts of a single population, and (b) the selection of the measurements in a systematic manner such that every population or part of a total population is represented in the sample by approximately the same number of measurements. The second characteristic is the *sine qua non*. *Systematic* selection of measurements from different populations for inclusion in a single sample leads to Lexian ratios less than 1; *random* selection of observations from either homogeneous or heterogeneous populations leads to Lexian ratios that are not significantly different from 1. Thus, if the balls in Culler's 10 urns that had different proportions of white and black balls (type *iii* sampling situation) were thrown in a single urn and random drawings were made from that urn, it is clear that the sampling situation would be no different from the one described as yielding a Bernouillian distribution of means with "normal" dispersion (type *i* sampling situation). The systematic selection of measurements from the different populations or different parts of the same population disturbs the operation of the normal law of error because this procedure introduces a negative correlation between the measurements included within the single sample: if one measurement is obtained from sub-population A, which has a true mean of 10, then another measurement must always be drawn from sub-population B, which has a true

mean of 15, *i.e.* one low value in a sample must always be matched by one high value (**146**, p. 348 ff.).

The condition of experimentation that yields Lexian ratios less than 1 has received less attention from investigators than the conditions that yield Lexian ratios greater than 1. In part this is attributable to the tendency to consider an underestimation of the reliability of experimental results as a commendable expression of the conservatism of science. Whether as a result of such fallacious logic, of a lack of belief in the validity of the proposition that underestimates occur when intra-sample negative correlations are present, or merely of the neglect of the problem, many of the experimental methods commonly used in the study of learning and memory are similar in structure to Culler's hypothetical urn-schema for demonstrating the conditions that produce underestimations of the reliability of means of single samples. In general, these methods involve either the systematic selection of subjects without taking this fact into account in estimating the error in the experimental results, or the systematic counterbalancing of the conditions under which the subjects work in an effort to avoid constant experimental errors such as would occur if the possible invalidating effects of practice, fatigue, quotidian variability, the intrinsic difficulty of the learning materials, etc., were ignored.

The thesis that such selective sampling as described by Culler leads to underestimates of the reliability of obtained means and differences between means is indubitably valid. Although there is no direct empirical evidence for such effects in the case of special methods for studying learning and memory, there is such empirical evidence in other instances, and the theoretical arguments are convincing. Fisher (**31**, pp. 71–72, 85–90) has shown that the systematic Latin squares used in agricultural experiments, *i.e.* the systematic arrangement of conditions in plots of a square field in the manner $\begin{matrix} ABC \\ BCA \\ CAB \end{matrix}$ rather than assigning the conditions to the sub-plots by chance, cannot lead to an accurate analysis of error, and frequently yields an overestimate of error when soil conditions are not constant throughout the field. Culler (**24**) found that some thoroughly practiced observers in lifted weight experiments have successively determined threshold values that vary less than the amount to be expected on the basis of chance ($L<1$), and found reason to believe that this occurred when the observers were using several differently sensitive criteria in making the judgments during each of the series of observations leading to a threshold determination.

Furthermore, the principle may be readily demonstrated by a simple statistical experiment with coins. The following experiment has been performed by the writer. Sixteen coins were tossed at one time and the number of "heads" were recorded. The tosses were repeated 25 times and the 25 records were taken as the first sample. Twenty such samples were obtained. The entire experiment was then repeated using first 12 coins and then 8 coins. Samples 1 through 20 obtained with the 16 coins were then combined with samples 1 through 20 obtained with the 12 coins. In this way 20 samples of 50 measurements that represented 2 essentially different distributions were obtained. The means, $\sigma_{dist.}$'s, and σ_{mean}'s of each of these 20 samples were calculated. Finally,

the standard deviation of the distribution of means from the 20 samples was determined. The mean of the means from 20 samples was 7.03 (true mean=7.00), with a standard deviation of 0.21. The average of the σ_{mean}'s computed from the $\sigma_{dist.}$ and N for each sample was 0.30. The Lexian ratio is 0.700, which is in accordance with expectations. The sampling procedure was repeated after combining the measurements obtained with the 12 and 8 coins. In this case the mean of the means from the 20 samples of 50 measurements was 5.07 (true mean=5.00) with a standard deviation of 0.16, and the average of the estimated σ_{mean}'s was 0.26. The Lexian ratio is 0.615. Finally, the measurements obtained with the 16 coins and with the 8 coins were combined to make 20 samples of 50 measurements. The mean and standard deviation of the means of the 20 samples were 6.07 and 0.13, respectively. The average of the estimated σ_{mean}'s was 0.37, and the Lexian ratio was 0.351.

The conclusive theoretical argument for the validity of the proposed relationship between the method of sampling and the inaccuracy in the estimation of the reliability of the mean has been given by Yule (**146**, pp. 285, 349). He shows that if every measurement obtained from sub-population A is a perfect representative of all the measurements in that population, and if every measurement from sub-population B is likewise a perfect representative of every measurement in that population, the standard deviation of means from successive samples must be zero, but the $\sigma_{dist.}$ of each sample is finite, and the estimated σ_{mean} is likewise greater than zero. For example, if 8 heads always turned up when 16 coins were tossed, and 6 heads always turned up when 12 coins were tossed, all of the means from successive samples of 50 tosses would be exactly 7.00, but the $\sigma_{dist.}$ for each sample would be 1.00 and the σ_{mean} calculated by the usual formula would be 0.14. Similarly, when an investigator selects 10 subjects for an experiment by taking one subject from each decile of a distribution of intelligence test scores, it is clear that from the standpoint of measurement in the experiment each subject represents a distinct sub-population of measurements, and that the estimated σ_{mean} of this group is an underestimate. This does not, however, hold true when the subjects are selected at random, even though each subject still represents a distinct sub-population of measurements, because the random selection of subjects will insure, on the average, the selection of more measurements from the region of the mean than from the extremes of the population of subjects. It is to guard against such a misinterpretation of the conditions that yield "undernormal" distributions of means that we previously inserted the clause that the number of measurements from the distinct sub-populations should be the same or approximately the same.

From the point of view of methodology, an important corollary of the general principle regarding the conditions of sampling that yield Lexian ratios less than 1 is that the degree of underestimation of the reliability of obtained means or differences between means depends on the difference between the true means of the sub-populations that have contributed to the sample. In the statistical experiment cited above the Lexian ratios for the 16–12 and 12–8 combinations were 0.700 and 0.615, whereas the Lexian ratio for the 16–8 combination was 0.351. The reason for the greater deviation from unity in the case of the 16–8 combination is evident. The absolute difference between the

means of the 2 sub-populations has no effect on the variation of the means obtained from successive samples, when the same number of measurements are obtained from each sub-population, but the $\sigma_{dist.}$ obtained for the entire sample must reflect the absolute difference between the means of the measurements from the different sub-populations. Accordingly, the $\sigma_{dist.}$ of the sample may be increased indefinitely without affecting the standard deviation of the obtained means from successive samples.

These principles regarding the presence and extent of underestimations of the reliability of experimental results permit the critical evaluation and comparison of many of the experimental methods commonly used in the study of learning without actually determining Lexian ratios, although comparisons in terms of inferences from these analytic principles lack the precision of comparisons in terms of the latter index. In the methodology of learning experiments the importance of these evaluative principles is particularly great, because investigators in the field have been prolific in the use of counterbalancing procedures for the control of potential experimental errors, such as practice, fatigue, quotidian variability, and differences in the intrinsic difficulty of various learning materials. This counterbalancing procedure has within it these essential conditions of systematic sampling that produce underestimates of the reliability of obtained means and differences between means. In some cases the application of these evaluative principles suggests a change in the method of statistical analysis of the data; in other cases it leads to a complete discrediting of the experimental method or to a limitation of the scope of its application.

Specific illustrations of the application of these principles and the types of methodological conclusions that result may serve a useful purpose in laying the foundation for later discussion of the indices of variable error (p. 348 ff.). (1) When an investigator needs to use more than one list of nonsense syllables in an experiment, he usually avoids constant experimental errors that might occur as a result of differences in the intrinsic difficulty of the lists used by using the lists an equal number of times under each of the experimental conditions. If, therefore, the lists actually vary in difficulty by significant amounts, the mean learning score for each experimental condition represents several essentially different sub-populations of measurements; the $\sigma_{dist.}$'s are higher than they should be if only one list were used; and the reliability of each of the means is underestimated. Furthermore, it is apparent that the underestimation of the reliability of each mean is proportional to the need for counterbalancing the lists in order to eliminate constant errors. If

the lists are equal in difficulty, there is no underestimation of the reliability of the mean, and the underestimation of the reliability of the mean increases as the actual difference in the intrinsic difficulty of the lists increases. An accurate estimate of the reliability of the means can be obtained either by comparing the experimental conditions separately for each specific list of nonsense syllables or by using standard lists of nonsense syllables that are known to be of approximately equal difficulty. Since many experimental studies of learning and memory involve too few observations for fractionation of the data in accordance with the first suggestion, the importance of experimental calibration of materials before use in experimental studies is again indicated. A third possibility in this and all other instances of systematic sampling is to replace the systematic counterbalancing procedure by a strictly random sampling procedure. Thus, the frequency of use of the various lists of nonsense syllables in the several conditions of the experiment could be determined by the throw of a die. This has the advantage that it enables an accurate estimation of the reliability of the obtained mean (31, pp. 85–90), but it has the disadvantage that it increases the actual unreliability of the mean. It may, however, be preferable in instances in which the first two possibilities cannot be realized.

(2) Another important instance of the underestimation of the reliability of experimental measurements occurs when the investigator fails to reduce practice and fatigue effects to a minimum before employing one of the several systematic counterbalancing procedures that have been devised (99).

Let us suppose that the practice effect in learning 12 nonsense syllables is such that subjects require, on the average, 16 trials to learn the first list and only 12 trials to learn the second list under the same experimental conditions, and that there is a commensurate practice effect when other basic experimental conditions are employed. In the simplest counterbalancing procedure the investigator eliminates the effect of practice as a source of error in his experiment by having one group of subjects learn under Condition X and then under Condition Y, while the second group of subjects learns first under Condition Y and then under Condition X. In computing the means, $\sigma_{dist.}$'s, and σ_{mean}'s for the 2 experimental conditions the investigator must combine the X measurements obtained at the 2 practice levels (and the Y measurements at the 2 practice levels) in order to achieve the equalization of practice. But, when this is done, the method is obviously similar to Culler's type *iii* sampling situation and to the sampling situation represented in our own coin-tossing experiment, and underestimation of the reliability of the means and of the differences between the means is to be expected. Furthermore, it should be noted that the estimate of the standard error of the mean difference between Conditions

X and Y by the formula that takes into account the use of the same subjects ($\sigma_{\text{mean diff.}} = \sqrt{\sigma^2_{\text{mean X}} + \sigma^2_{\text{mean Y}} - 2r\sigma_{\text{mean X}}\sigma_{\text{mean Y}}}$) yields a spuriously high estimate of error because the r between the X and Y measurements is attenuated when the conditions are counterbalanced to control for practice.

The point of greatest methodological interest is that the underestimation of the reliability of the individual means and of the difference between the means is a function of the gross practice effect present in the experiment. If the subjects have been brought close to a practice level before the experiment is begun, the underestimation is slight (the attenuation of the intercorrelation of the X and Y measurements is likewise slight). The efficacy of the counterbalancing procedure from this point of view therefore depends on the type of material being learned, the amount of preliminary practice given to the subjects, etc.

It may appear that the criterion that refers to the accuracy of the statistical estimates of error is applicable only to the evaluation of different methods for the control of potential experimental errors such as practice, the difficulty of the learning material, etc. This is not a valid inference, because the evaluation of these methods of experimental control is so dependent on the mean difference between the sub-populations included within a single sample, and this difference is clearly a function of the method of experimentation used (*i.e.* whether the anticipation method, method of complete presentation, etc.), on the nature of the material used in the experiment, and on the type of measures of learning obtained. If, for example, it can be shown that lists of nonsense syllables differ more in difficulty than lists of words, this fact is not only of importance in predicting the amount of variable error that will be involved in experiments in which those 2 materials are used, but it likewise indicates that the use of unselected lists of nonsense syllables in a counterbalanced order will, on the average, lead to greater inaccuracy of the statistical estimates of error in terms of which the reliability of experimental results are to be evaluated.

B. The Amount of Variable Error as a Criterion for Evaluating Experimental Methods

There remains the problem of determining the specific indices in terms of which the amount of variable error resulting from the use of different experimental methods may be compared. The apparent simplicity of this problem is deceptive. Several different indices have been proposed, and the only one that has received extensive use up to the present time—the reliability coefficient—has yet to be operationally defined in a manner acceptable to all students of the methodology of learning.

The major source of the difficulties encountered in the attempt to measure the variable error that results from the use of different experimental methods, materials, and measures is the occurrence of non-random variations between successive measurements. These invalidate attempts to estimate the true dependability of experimental

measurements from the known variability of successive measurements. In short, it has been found difficult to obtain indices of variable error that are related to the actual variability of means from successive samples in such a way that the Lexian ratio is not significantly different from unity. Whenever the variable error involved in successive measurements is represented by any one of the indices of reliability and these indices are used to compare experimental methods, it is clear from what has been previously said regarding the Lexian ratio that the latter must be unity in the case of the measurements obtained by each experimental method or must differ from unity by the same amount and in the same direction in the case of the measurements obtained by each of the experimental methods. If this condition does not exist, the true relative reliability of the measurements obtained with different experimental methods cannot be determined from indices of the variability of obtained measurements.

The three indices of reliability that have been proposed are: (a) the reliability coefficient, or the correlation between successive measurements obtained from the same subjects under the same experimental conditions; (b) the actual variability of successive measurements on the same subject or with a group of subjects under the same experimental conditions as represented in the standard deviation of the distribution of obtained measurements; and (c) the actual variability of means obtained from successive groups of measurements under the same experimental conditions. These indices have as their immediate objective either the determination of those experimental methods that give the closest approximation to the true measurements for a given subject—the primary concern of those interested in measuring individual differences; or the determination of those methods that yield a mean value of a series of observations on the same or different subjects which is nearest the true mean value for the subject or population sampled and for the given constant conditions of experimentation—the primary concern of those interested in determining the effect of experimental variables other than the subject. The proposed indices of variable error reflect this somewhat divergent interest, but any index must be valid for the evaluation of methods used in both types of studies or adequate for neither. The experimental method that gives the most reliable determination of a mean value must have given, on the average, the most reliable determination of each individual measurement summarized in the mean.

1. *The Correlation of Repeated Measurements on the Same Subjects as an Index of the Reliability of the Methods, Materials, and Measures Used.* The variable error involved in the use of different methods, materials, and measures in the study of learning is most frequently measured in terms of the reliability coefficient. This apparent preference of investigators is attributable to several factors. (a) Since the introduction of the reliability coefficient by Spearman in 1904 (**108**) it has been used to the exclusion of other indices in the evaluation of mental tests. (b) The first study of the reliability of single measurements obtained with animal and human mazes used this index (**49**). (c) The reliability coefficient is needed as a correction for the attenuation of r's obtained in the study of the community of mental functions that are measured by different methods, materials, and measures (see p. 327). (d) The reliability coefficient is independent of the units of measurement employed, and is for this reason applicable to the comparison of alternative forms of any aspect of the methodology of learning experiments. Thus, it may be used to compare the reliability of different methods or materials when the difference between the methods or materials occasions a mean difference in the learning scores obtained, and it may be used to compare the reliability of different measures of learning and retention, such as trial scores, error scores, and time scores. (e) The reliability coefficient measures directly the relationship in which the experimenter is most interested, namely, the ratio of the amount of variable error involved in individual measurements to the amount of difference to be expected between the means of different groups of measurements.

When a reliability coefficient is obtained under conditions that satisfy the assumptions involved in its computation, it represents the ratio between the amount of variance that is attributable to true differences in the abilities of the subjects included in the group (σ^2_{true}) and the total variance of the obtained measurements ($\sigma^2_{dist.}$), *i.e.* $r=\sigma^2_{true}/\sigma^2_{dist.}$ (**25, 129**). Therefore, $(1-r)$ is a ratio of the amount of variance attributable to chance errors of measurement ($\sigma^2_{meas.}$) and the total variance of the measurements ($\sigma^2_{dist.}$). Stated in terms of variability rather than variance, the $\sqrt{1-r}$ is the ratio of the standard error of measurement ($\sigma_{meas.}$)—the root mean square deviation of the second measurement from the first measurement for each subject—to the standard deviation of the scores obtained from the group of subjects used ($\sigma_{dist.}$), and this in turn may be interpreted as the per cent of the total variability that is occasioned by chance errors of measurement.

From the point of view of mathematical adequacy, the reliability coefficient is an ideal index of the dependability of the measurements

obtained with different experimental methods. If the prerequisites for the comparison of the measurements obtained with different methods have been satisfied, it is clear that the method which yields the highest reliability coefficient is the one that gives the finest differentiation of the abilities of different subjects. Furthermore, it is legitimate to conclude that the method that yields the highest reliability coefficient also permits the investigator to make finer differentiations between the effects of experimental variables other than the ability of the subject, since there is no reason to assume that the true differences between subjects and true differences between the performances of the same subject or group of subjects under two experimental conditions belong to different categories. The reliability coefficient is, therefore, an appropriate index for use in evaluating methods that are to be used in the study of either individual differences or group differences. In both instances it measures the probable ratio between the amount of variation in measurements that is to be expected as a result of the intrusion of chance factors and the amount of variation that is to be expected as the result of true differences between the effects of experimental variables.

However, the prerequisites for the valid use of a correlation between two series of measurements as the reliability coefficient are difficult, if not impossible, to fulfil when the measurements represent mental functions. The first prerequisite is that the two series of measurements must represent the same true abilities of the subjects, *i.e.* they must be measures of the same thing. This follows from the fact that the reliability coefficient varies as the σ_{true} in the numerator of the ratio $\sigma_{true}/\sigma_{dist.}$ varies; and if the two series of measurements represent essentially different mental functions, the σ_{true} represents only the common factor in the two measurements, and the "true" or non-accidental determinants of the two sets of measurements are treated as though they were accidental. The second prerequisite of a valid reliability coefficient is that the accidental errors in the two series of measurements must not be correlated. If the deviation from the subject's true score in the first test is correlated with the deviation from the subject's true score in the second test, it is clear that the common element in the two errors is included as a portion of σ_{true} and leads to a spuriously high reliability coefficient. The third prerequisite is that the accidental errors in the first series of measurements must not be correlated with either the true scores in the first series or the true scores in the second series. Similarly, the acci-

dental errors in the second series of measurements must not be correlated with the true scores in either the first or second series of measurements. Finally, the reliability coefficient is a valid index of the true variable error only when the measurements in the two series are normally distributed, or at least have not been artificially restricted in range.

These prerequisites are so restrictive that it is doubtful whether a completely valid estimate of the ratio between accidental errors and true differences between experimental conditions (subjects) can ever be obtained in the case of mental measurements. Nevertheless, it should be possible to obtain reliability coefficients for different experimental methods that are adequate for use in comparing the dependability of the measurements obtained with those methods, provided the abrogation of the prerequisites is a constant. The problem therefore becomes one of stating the operations to be performed in the determination of reliability coefficients so that the reliability coefficients are maximally and equally accurate for all of the methods, materials, or measures that are to be compared. In order to do this it is necessary to consider the specific characteristics of the mental functions that are to be measured, since these specific characteristics may be sources of error in the determination of the coefficients. The search for some one most valid method for determining the reliability coefficient in the case of mental functions must be deprecated. Thus, specific methods for determining the reliability coefficient used by students of sub-human learning may be valid for the types of experiments performed with such subjects but less acceptable than other methods for use with human learning. For example, rats are usually given one or two trials per day, whereas human subjects learn mazes and other materials under conditions of massed practice. Consequently, any reliability coefficient based on the scores obtained on different trials must involve quite different amounts of correlation between errors of measurement, quite different degrees of attenuation as a result of the quotidian variability of the subjects, etc., when used with rats and with human beings. The proper inference is, of course, that a single set of experimental operations need not be accepted as the most valid for use in comparing methods, materials, and measures employed in the study of verbal learning merely because it is considered most valid in the case of human motor learning or the learning of rats.

In view of this intimate relationship between the adequacy of a particular method for determining the reliability of experimental

methods used in the study of learning and the particular characteristics of the learning process, it is profitable to review the factors that condition the size of the reliability coefficient in learning experiments before considering the several particular methods that have been proposed for determining the coefficient. In considering these factors two questions are paramount: (1) How does the factor affect the size of a reliability coefficient? (2) How does the factor affect the validity of comparisons of reliability coefficients that have been determined for different methods, materials, and measures used in the study of learning?

(a) *Legitimate Accidental Errors,* i.e. *Errors of Measurement.* In so far as the measurements obtained in a learning or memory experiment have been determined by chance, accidental, or sporadic factors in the experimental situation the reliability coefficient should be decreased in size, because it is these chance factors that cause the variation in successive experimental measurements that the student of learning wishes to reduce or eliminate in his experiments. Therefore, whether a reliability coefficient is accepted as a valid index of the reliability of a particular experimental method, material, or measure depends to a large extent on the definition of the variable determinants of performance that the investigator intends the reliability coefficient to reflect. This, in turn, depends on the definition of the "true" measurements from which the accidental deviations are measured.

There is a sharp difference of opinion regarding these definitions. Cureton (**25**), for example, contends that the "true" ability of the subject is, in effect, his mean performance in an infinite number of independent measurements with the same or comparable tests. Therefore, two sets of errors taken together constitute the errors of measurement: the *response errors, i.e.* the day-to-day and hour-to-hour (intrinsic) variability in the performance of the subject, and the *test errors, i.e.* the failure of the test to sample adequately the trait under consideration. Since the day-to-day variability of the subject is considered as a legitimate error of measurement in estimating the reliability of measurements, it follows that the measurements used to determine reliability coefficients should be obtained on different days in order to avoid a correlation of errors. Comparable definitions of the "true" score and of the errors of measurement have been stated or implied by Kelley (**58**, p. 200), Paterson *et al.* (**88**, pp. 26–28), and by Spence (**110**) and Leeper (**64**) in their critiques of the methods used to determine the reliability of mazes. In fact, all who have used correlations between measurements obtained on different days as reliability coefficients have assumed that the day-to-day variability of the subject should be considered as an error of measurement.

An opposed point of view, especially represented in a recent paper by Anastasi (**4**), and in earlier criticisms of Spearman's definition of reliability

by Brown (12, 13), is that the reliability coefficient for a test should not reflect the intrinsic variability of the trait as represented in the day-to-day variations in performance, but should reflect only test errors. The argument is that the test should not be penalized for measuring an intrinsically variable trait. Thus, Anastasi (4, p. 322) says, "It seems somewhat paradoxical to label a test unreliable simply because it may be a very sensitive measure of a phenomenon exhibiting marked daily variations." Brown (13) gave an additional cogent argument for eliminating intrinsic variability from consideration when he showed that the amount of variability was correlated with the quantity of the ability possessed by the subject. This condition violates a major statistical assumption involved in the definition of the reliability coefficient. The recommendation is that the reliability of a test be determined from two series of measurements obtained during the same experimental period; otherwise, the coefficients are too low and the extent of the false attenuation varies from test to test, thus vitiating comparisons. This limitation of the variable errors that are considered errors of measurement is accompanied by a definition of the "true" ability of the subject as his ability at the time of testing, including the effects of all the mental and physical influences that affect his performance at that time.

This dichotomy of variable factors rests on evidence that the measurements obtained from subjects on the same day differ less than measurements obtained on different days. Thus, Woodrow (143) has demonstrated that the differences between the mean performances of subjects in simple tasks on different days are greater than is to be expected as a consequence of chance combinations of the variable factors that produce variations in performance within a single day, and has named this day-to-day variability *quotidian variability*. Likewise, there have been several reports of lower r's between measurements obtained on different days than between measurements obtained during the same experimental period (89, 116), and the validity of this generalization has been assumed by several writers (4, 25, 27). However, there has been no effective analysis of the nature of the factors that produce quotidian variability, and one may question whether it is a general factor, *i.e.* of such a nature that it depresses or enlivens all mental functions at the same time, and whether it is either necessary or of great importance. Regarding the former, Hollingworth (44) failed to show a correlation between the *variations* of simple mental functions, such as cancellation and color naming, under conditions that should have produced a correlation if quotidian variability were a general factor. Regarding the inevitability and importance of such variations, it is of some significance that Woodyard (144), in the case of simple and complex mental tests that were chiefly non-learning in type, found that the time interval between measurements had only a slight relation, if any, to the size of the r's obtained. Studies of quotidian variability in the case of the more complex learning tasks such as used in the laboratory have not been made.

The arguments for the elimination of intrinsic variability from consideration as a source of errors of measurement appear to be valid, but the mode of elimination proposed—the use of measurements from the same experimental period—cannot be accepted as a general principle in the study of memory and learning methods. In

the first place, one may question whether the elimination of intrinsic variability in reliability determinations has the importance that has been ascribed to it. The inclusion of quotidian variability can result in erroneous comparisons of the reliability of different tests only in case the basic abilities measured by the tests are significantly different, but reliability comparisons have practical significance for the student of learning methods only when the tests or methods are known to measure approximately the same ability. Comparisons of the reliability of intelligence tests, rat mazes, stylus mazes, and nonsense syllables have no practical value and usually have no meaning even though quotidian variability has been eliminated in each case. In short, the fact that reliability coefficients reflect quotidian variability does not vitiate most of the comparisons that use them. Furthermore, it is questionable whether one should always eliminate intrinsic variability in computing reliability coefficients that are to be used in correcting intercorrelations of different abilities for errors of measurement, since most studies of the abilities represented in different tests of learning require the learning of different tasks on different days. A correction for quotidian variability may be advisable.

Second, the determination of both measurements during the same experimental period frequently is either impossible or introduces serious systematic errors when learning and memory methods are involved. For example, the specific positive and negative transfer of learning from task to task, and warming-up and fatigue may be introduced or accentuated. These may be partially eliminated or counterbalanced when the simpler learning and memory tests are involved, but the procedure becomes impossible in the case of the important tests such as the maze, rote memorization to complete mastery, etc. By "impossible" is meant that the complete mastery of several mazes, etc., is impossible. The use of alternate trials during the learning of a single material is, of course, possible; but there are other objections to this procedure.

The most important objection to the intra-day reliability coefficient is that another presumably important source of variable error in experiments on learning is eliminated from consideration when the intrinsic variability is eliminated. As previously indicated, no adequate analysis of quotidian variability has been made, yet it is fairly certain that quotidian variability cannot be considered as equivalent to intrinsic variability, when the latter signifies the variability of measurements occasioned by extra-laboratory circumstances. The variation in measurements from day to day is probably caused in part by actual variations in the experimental situation, *e.g.* the subtle changes in the relation-

ship between the subject and the experimenter, misunderstanding of instructions, errors in the recording of the subject's responses, changes in the general conditions of light and noise. Such variations occur; they are frequently of such a type as to influence the performance of the subject throughout large sections or the whole of the experimental period; and they are an important source of unreliability in determinations of experimental differences. If, therefore, the correlated measures are obtained from the same experimental period, the *r*'s are too high because these variations in experimental conditions affect both measures equally. The use of measures obtained on different days is, of course, no sure method for eliminating the correlation of such errors, but the probability is decreased. Leeper (64) has given an extended analysis of these "systematic" errors in the case of rat learning.

The practical significance of eliminating correlations of these situation errors in the evaluation of learning methods is that the frequency and magnitude of the effects of these disturbances may be a function of the type of material being learned, the method of experimental control of the learning process, and the measure of learning used.

From these considerations, it appears that there are at least two legitimate errors of measurement, neither of which should be permitted to operate during both measurements of the performance of a single subject: (1) errors of measurement attributable to the inconstancy of certain aspects of the experimental situation, and (2) errors of measurement attributable to the peculiar characteristics of the test or learning material used, *i.e. test errors.* Furthermore, the intrinsic variability of the subject may require elimination in special instances, but such elimination probably should not be achieved by having both measurements obtained on a single day.

(b) *Errors Involved in the Use of Test-Retest Reliability Coefficients.* The strict interpretation of the reliability coefficient requires that the subjects be retested with the same material, *e.g.* the same list of nonsense syllables or the same maze, in order to reveal the *test errors* specific to that material.

What makes a learning material unreliable, aside from its differential susceptibility to situation errors, has not been treated in detail by students of learning, but Willoughby (139) has given an illuminating analysis in the case of non-learning tests that may be applied to learning materials. When a list of nonsense syllables that includes DEQ is presented to a subject for the first time he may happen to associate DEQ with DECK and learn the list especially rapidly for this reason. But, if the subject were returned to the original naïve state and the list of syllables were presented again, the syllable DEQ might on this occasion arouse the associate DICK and hinder learning by causing the nonsense syllable to be falsely learned as DIQ. The difference between the associations aroused is the essence of the *test error* and the test is unreliable in proportion to the frequency and magnitude of the effects of such differences. It is clear that the reliability of the item DEQ can be determined only by

repeating the list in which DEQ occurs. However, this procedure introduces a number of factors that vitiate the "reliability" coefficient so obtained, because the learning of the list on the second occasion is not uninfluenced by the previous learning.

Complete amnesia for the first learning can rarely, if ever, be obtained in the study of learning materials, although it may be approximated in the case of mental tests of other types (**25, 88**). In so far as complete amnesia is not achieved, the correlation between the measurements obtained may violate all but one or two of the statistical assumptions involved in the use of that r as a measure of reliability. (1) The abilities measured in the 2 tests are not the same. Thus, the correlation represents either the relation between a test of learning and a test of retention plus learning, or it represents the correlation between what may be essentially different stages of the learning process. Although the relatively low correlation between the speed of learning and the amount retained cannot be used as an argument for the distinction between a learning ability and a retention ability, since this low correlation may merely reflect the unreliability of the measurements, the distinction is probably valid on other grounds. Likewise, there is conclusive evidence that the different stages of many learning processes, such as are represented in correlations of the measurements obtained during the first and second halves of the period required for mastery of nonsense syllables, mazes, etc., are not equally representative of a unitary process. For example, intra-serial interference effects undoubtedly occur during serial learning, as suggested by Foucault (**32**) and there is ample evidence, at least in the case of nonsense syllables and mazes, that these effects change from trial to trial during learning as a concomitant of changes in the degree of learning of the interfering items (**69, 70, 71, 72, 86**). Therefore, such correlation coefficients may represent the relation between the ability to learn when inhibition from other learning is minimal and the ability to learn in the face of marked inhibition.

(2) Such positive transfer of learning from the first to the second test is probably not constant in amount, and it may be, under some circumstances, correlated with the test errors involved in the first test. Thus, a subject's second score may be much better than it should be merely because DEQ was learned as DIQ and the time spent in eliminating this error during the first test led to a great degree of overlearning of all the other syllables in the list. Obviously, the effect of such intercorrelations on the size of the reliability coefficient cannot be phrased in terms of a general prediction, but it is generally assumed that an attenuation of the r is the most common consequence.

(3) The transfer of learning may result in a correlation between the test errors. Thus, in the example previously cited, the subject may respond to DEQ the second time it is presented by again thinking of DECK, in which case his speed of learning is beneficially affected by the same chance test error in both instances. (4) Finally, the use of measurements obtained from the learning of the same material violates the principle (**25**) that the series of measurements that are correlated to obtain the reliability coefficient should be statistically comparable, *i.e.* equally reliable, and with approximately the same σ's. It is known from the work of Anastasi (**4**) and others that the reliability

of measurements obtained during learning increases with the increase in level of learning achieved, and that the $\sigma_{dist.}$ likewise changes.

The only effective method for eliminating these difficulties without relinquishing the ideal of retesting the subject on the same material involves the introduction of long periods of rest between tests in the hope that the effect of the first test will be obliterated. This method may be satisfactory in the case of non-learning tests, as maintained by Paterson *et al.*(**88**), and Cureton (**25**), provided the fundamental ability being tested has not undergone growth or decay by the time the second test is made. However, there is an important difference between the usual mental test in which the subject reacts to an individual item for only a few seconds without the intent to learn, and the test of learning in which the subject persists in his efforts to master a relatively small number of items. Thus the method is probably satisfactory for tests of immediate memory, at least in so far as specific positive and negative transfer effects are concerned, but is invalid for measuring the reliability of tests of complex serial learning unless extremely long rest periods are introduced in studies with mature subjects, *i.e.* subjects in whom the ability is not undergoing marked progressive change. It is well known that most laboratory learning is relatively resistant to complete forgetting, at least when retention is measured by the saving method. As an example, in the correlation studies of Hunter and Randolph (**52**) with stylus mazes and nonsense syllables, the subjects showed an appreciable retention after intervals as great as 160 and 50 days, respectively, even though the first tests consisted of only 6 trials. These restrictions obviously extend to the comparison of reliability coefficients obtained with different learning materials even though a strictly valid reliability coefficient is not demanded. Any comparison of learning and retention scores assumes that the degree of retention is equal for the various materials compared. This assumption can rarely be justified.

(c) *Comparable Materials.* A second solution to the difficulties involved in the use of the test-retest method involves the re-definition of *test errors.* In this method no attempt is made to determine the reliability of a particular list of nonsense syllables or a maze; instead, one attempts to determine the extent to which apparently similar materials measure the same ability. Thus, Kelley (**58**, p. 200) defines a "true" score not as the average score on an infinite number of repetitions of a particular test, but as "the average score on an infinite number of strictly comparable tests." As has been noted by many writers (**26, 63, 119, 139**) this definition of reliability states the limiting case of measurements of validity. The mode of testing the reliability of mazes and the reliability of nonsense syllables by this method therefore involves the correlation of measurements obtained on "strictly comparable" mazes and "strictly comparable" lists of nonsense syllables.

Comparable tests have been defined by Kelley (**58**, p. 203) as tests in which "(1) sufficient fore-exercise [has been] provided to establish an attitude or set,

thus lessening the likelihood of the second test being different from the first, due to a new level of familiarity with the mechanical features, etc.; (2) the elements of the first test [are] as similar in difficulty and type to those in the second, pair for pair, as possible; but (3) [are] not so identical in word or form as to commonly lead to a memory transfer or correlation between errors." In short, comparable tests should measure the same ability in the same units with the same error, and should therefore have equal standard deviations. If these conditions exist, the tests have equal reliability, and the correlation between the test scores represents the reliability of either test. Cureton (**25**) has recently suggested that the reliability of each test may be determined separately by using 3 comparable forms and by then evaluating the reliability of each form by means of the Spearman formula for the correlation between a single test score and "g," namely, $R_1=r_{11}r_{11}/r_{11}$. Dunlap (**27**) has shown that the use of this method assumes that the mean tetrads are equal to zero, *i.e.* that there is only one group factor represented in the tests and that the specifics are entirely errors of measurement. Thus, the use of 3 comparable forms permits a check on the extent to which the scores on the individual forms are representative of the ability measured by all three, and has wide applicability in studies of the methods and materials used in the study of learning and memory.

Attempts to fulfil the requirements of comparable tests in the case of learning materials encounter difficulties even greater than those encountered in the case of non-learning tests. These difficulties have undoubtedly served to discourage the use of this method for determining the reliability of learning materials. In the first place, the changes due to habituation to the task of learning are probably greater in amount and are more difficult to eliminate than in the case of non-learning tests. In the second place, the selection of materials that are "similar in difficulty and type . . . yet not so identical in word or form as to commonly lead to memory transference" (**58**, p. 203) is practically impossible in view of the ubiquity of specific positive and negative proactive transfer effects in learning. Is it possible to obtain 2 lists of nonsense syllables that are sufficiently different in form to be free from specific proactive transfer? At present, our knowledge of proactive facilitation and inhibition is not sufficiently precise to define the materials or conditions of experimentation that would successfully minimize these disturbing effects. It is, however, known that the formal identity of the materials (**15**, **23**), the time interval between the learning of 2 materials (**74**), and the degree of learning of the first material (**15**, **59**, **106**) condition the amount of specific transfer between non-identical materials. These factors must be considered in evaluating reliability coefficients for learning materials that have been obtained by means of "comparable" samples of the materials. And in comparing reliability coefficients obtained with different materials there is the additional important question whether the proactive transfer effects have been equivalent for all the materials.

(d) *Progressive Errors,* i.e. *Practice and Fatigue.* The matter of progressive improvement and progressive decline in the efficiency of performance from test to test has been a persistent problem in reliability measurements since the reliability coefficient was first introduced (**12**, **13**, **109**, **140**), although it has received very little attention

in recent reviews (**64, 110, 130**) of the methods used for the determination of the reliability of learning materials and methods. Since fatigue effects are so readily eliminated in most studies of learning by the proper spacing of the work periods, we may confine the discussion largely to the so-called general "practice effects." A distinction between general habituation to the task of learning and specific positive and negative transfer is implied. Among the factors that contribute to increases in the efficiency of performance from one period of learning to the next are changes in the general attitude of the subject toward the task of working under laboratory restrictions, changes in his attitude toward the experimenter, changes in the method of learning used and in the understanding of the problem, and changes in the frequency of disturbing emotional responses.

When all subjects take the 2 forms of a test in the same order, the reliability coefficient is unaffected if the general improvement is constant for all subjects; the r's are spuriously high if the improvement is positively correlated with the speed of learning in the first test; the r's are attenuated if the improvement varies from subject to subject but is uncorrelated with the speed of learning in the first test. It is probable that the effect of this factor must be estimated independently for each experiment.

The experimental precautions that may be used to eliminate or minimize the effects due to progressive habituation are not altogether satisfactory. The best that can be expected is a reduction in the amount of error. Some investigators, following the suggestion made by Kelley (**58**, p. 203), have used a short preliminary sample of the task. Thus, Spence's subjects (**110**) learned a three-alley stylus maze before learning 2 more complicated mazes that were to be used for reliability determinations, and Garrison (**36**) gave his subjects preliminary training on a three-letter Peterson Rational Learning Problem before having them learn longer problems. It is, however, doubtful whether a very short preliminary sample of a task provides sufficient practice to eliminate even the initial marked changes in performance. McGeoch and Oberschelp (**84**, p. 164) have shown "that, with the 6- and 12-letter [Peterson Rational Learning] problems, ease of learning is greatly increased by practice at other problems and that practice on two problems [6-, 12-, or 18-letter] yields a distinctly greater increase than does practice on one." Garrison (**36**) likewise presents evidence for a considerable change in performance from the first to the second of 2 eight-letter rational learning problems, even though the preliminary problem had been learned. In the case of stylus maze learning there is no conclusive evidence that practice effects persist throughout the learning of a number of mazes, but the data presented by Heron (**40**), Spence (**110**), and McGinnis (**85**) suggest that a considerable amount of preliminary training must be given before the possibility of further habituation to the task no longer exists. In the complete memorization of lists of nonsense syllables Luh (**68**) and Ward (**137**) have shown that naïve subjects must learn at least 5 lists before they reach even an approximately constant level of performance, and McGeoch (**82**) has found a similar persistence of practice

effects in the memorization of lists of 10 adjectives. Even in the case of the memory span, changes in performance attributable to practice are appreciable long after the subjects begin the experiment (**73**).

Another technique that has been used in an effort to reduce the magnitude of changes in general habituation and of specific positive and negative transfer from task to task involves the introduction of long periods of rest between the learning of the different tasks. Thus, Heron's subjects (**40**) learned 5 stylus mazes one week apart. Hall (**38**) likewise employed a one-week interval in his study, and the implication is that this interval reduces both the general practice effect and the specific proactive transfer by an appreciable amount. It is, however, doubtful whether the general habituation to the task of learning is forgotten with sufficient rapidity to permit a significant reduction of this source of error by the use of a one-week interval between tasks. There is no conclusive evidence on this point. However, in the case of the stylus maze Tsai (**133**) found that the percentages of saving in errors in the relearning of an irregular maze after 1, 2, 3, 5, 7, and 9 weeks were 94, 90, 85, 86, 84, and 81, respectively. The savings in trials and time were correspondingly high. If it is assumed that the transfer of training from maze to maze is roughly proportional to the degree of retention of the first maze, the need for more extended periods of rest between successive mazes is indicated. The same is probably true of many other types of learning. For one thing, it is conceivable that learning how to learn various materials and learning how to learn under laboratory conditions may have the resistance to forgetting that characterizes meaningful materials or greatly overlearned materials or acts.

A third method for eliminating the effect of practice (or fatigue) is to measure each individual several times and then divide his measurements into 2 groups in such a way that practice produces no mean difference between the averages of the groups. The correlation between the averages of the groups is then used as the reliability coefficient. Several methods for counterbalancing practice have been used. Spearman (**109**, p. 274) recommended that "a test of verbal memory, for instance, might well consist of memorizing twenty series of words (exclusive of some preliminary series for 'warming up'). Then series 1, 3, 5 . . . 19 would suitably furnish one group, while the even numbers gave the other. Any discrepancy between the averages of the two groups, might, as a rule, be regarded as practically all due to the 'accidents.'" It is, however, apparent that the grouping of odd and even measurements will not counterbalance the effects of practice; some system such as the *ABBA* is needed. Even this is not entirely satisfactory unless the measurements are obtained after a number of preliminary practice periods, because the counterbalancing procedure assumes that the practice curve is linear, and this is not true, at least in the case of verbal learning, until the fourth or fifth list has been learned (**26, 68, 104, 137**). Furthermore, the very rapid non-linear drop in the practice curve lasts throughout a greater number of lists in the case of meaningless materials than in the case of meaningful materials (**26, 104**), and reliability comparisons of these materials may be vitiated by this fact whenever preliminary practice is not given before the counterbalancing procedure is used. Presumably, similar differences occur with other learning materials. In studies in which fatigue, as well as practice, may be operative, it is necessary to use

special counterbalancing procedures. Thus, Anastasi (2, 3, 4) and others have made a practice of grouping odd and even measurements obtained during a single experimental session.

Special difficulties are encountered in attempts to control practice in methodological studies in which an attempt is made to compare the reliability of several materials or methods by administering them to the same subjects. An example is presented by the study of Stroud, Lehman, and McCue (115) in which a comparison was made between the reliability of lists of 6, 12, and 18 nonsense syllables by correlating scores on comparable lists. Twenty-five subjects learned the lists in the order 6, 6, 12, 12, 18, 18; 25 learned in the order 12, 12, 18, 18, 6, 6; and 26 learned in the order 18, 18, 6, 6, 12, 12. In computing the reliability coefficients the records of all 76 subjects were combined. This is obviously an application of the familiar systematic counterbalancing procedure in which different groups of subjects go through the experimental conditions in the order *ABC, BCA,* or *CAB,* where *A* represents both forms of the test to be used in determining the reliability coefficient, or where *A* represents a single series of measurements that is to be split into 2 parts in order to determine the reliability coefficient (*e.g.* as in the correlation of errors on odd and even trials, or errors in the first and second half of learning, or measurements obtained during learning and relearning).

Although this method may be satisfactory for eliminating practice or fatigue as determinants of the average scores for measurements obtained with the different methods and materials, the effect of the procedure on the reliability coefficients is to make them uniformly higher than they should be. When different subjects in the major group learn the same material at different stages of practice the sample is heterogeneous with respect to practice, and this factor is a common determinant of all the measures obtained simultaneously or in immediate succession. For example, individuals *X* and *Y* may be equal in ability on task *A,* but *X* requires 10 trials and 8 trials in the mastery of the 2 forms of *A* because it is the first task to be encountered; whereas, individual *Y* requires 6 trials and 5 trials to master the same forms of task *A* because it is learned after having had experience with tasks *B* and *C*.

This is merely a special case of a more general principle regarding the effect of this counterbalancing procedure on intercorrelations between the measurements obtained. The intercorrelations of the measurements obtained for the counterbalanced tasks may be either spuriously high or attenuated. If only 2 tasks are used and they are given to some subjects in the order *AB* and to others in the order *BA,* the investigator introduces a negative correlation between general improvement and the ability to learn specific tasks. Thus, individuals *X* and *Y* may be equal in ability on tasks *A* and *B,* but *X* scores 10 on *A* and 8 on *B* because he learns them in the order *BA*. Consequently, the *r* between scores on *A* and scores on *B* is attenuated. The effect of the counterbalancing procedure is, however, much more complicated when more than 2 tasks are used. If 3 or more tasks are used, the intertask *r*'s are always maximally attenuated when every possible serial order of the task is used, *e.g. ABC, BCA, CAB, ACB, BAC, CBA.* But, few investigators attempt to counterbalance the positions of the tasks in the practice series so completely. They use, for example, only the simple rotation *ABC, BCA, CAB,* or *ABCD,*

BCDA, CDAB, DABC. When this is done the correlations may be either spuriously high or attenuated, depending on the number of tasks used. A spurious positive correlation between the measurements obtained with any 2 tasks is introduced whenever those tasks are paired in the same order at different points in the practice series. There is, in this case, a constant practice difference between the subjects that is in the same direction for both tasks. For example, in the practice order *ABC, BCA, CAB,* one subject learns *A* and *B* early in practice and another learns *A* and *B* late in practice, thus introducing a spurious correlation due to the heterogeneity of the population. But, since a third subject learns *B* early in practice and *A* late in practice, an attenuating influence is introduced; hence, the indeterminateness of the effect. Obviously, if the number of tasks is increased and the simple rotation method is followed, the possibility that the intercorrelations will be spuriously high increases, because tasks *A* and *B* occur in sequence at different stages of practice for one additional subject or group of subjects every time the number of conditions or tasks is increased by one, but task *B* occurs early in practice and task *A* occurs late in practice for only one subject or group of subjects regardless of the number of conditions or tasks rotated.

When the practice effect from task to task is not equal for all materials, and it is rarely so when the materials are of different intrinsic difficulty, the reliability coefficients obtained by these counterbalancing procedures are no more comparable than those obtained from groups of subjects that include different ranges of talent, and a control group procedure is preferable. Nevertheless, such comparisons may be valid if the subjects have been extensively practiced in all the tasks before the observations are made.

(e) *The Subjects Used.* Since the reliability coefficient is a ratio between σ_{true} and $\sigma_{dist.}$, the range of learning or memory ability present in the particular group of subjects used is one determinant of the magnitude of the coefficient obtained. That is, if an investigator reports a reliability coefficient of 0.50 for a particular memory material, it cannot be assumed that this coefficient is a general index of the reliability with which individual differences may be measured with that material. This point has been clearly stated by Kelley (**57**); and Tolman and Nyswander (**123**), Tryon (**130**), Spence (**110**), and Leeper (**64**) have emphasized the importance of considering the range of talent in evaluating reliability coefficients obtained in studies of learning. Since there is, in general, a positive correlation between the reliability coefficient and the range of ability represented in a group of subjects, it cannot be concluded that the maze used by one investigator is more reliable than the maze used by a second investigator merely because the first reports a reliability coefficient of 0.75 and the second reports a coefficient of 0.50.

It has been suggested that inter-experiment comparisons of reliability determinations may be made if the investigators report $\sigma_{meas.}$'s or $\sigma_{dist.}$'s (**57**; **58**, p. 221). Using these measures, Kelley has developed a formula for adjusting

coefficients for varying ranges of talent. The formula is $\sigma/\Sigma = \sqrt{1-R}/\sqrt{1-r}$, where σ is the obtained $\sigma_{dist.}$, r is the obtained reliability coefficient, Σ is the $\sigma_{dist.}$ of the adjusted range of talent, and R is the predicted correlation coefficient for the new range of talent. The basic assumption involved in the use of this formula is the constancy of the standard error of measurement throughout the entire range of talent.

The usefulness of this formula in comparing the reliability coefficients obtained with different learning methods is questionable. In the first place, the proof of the formula is open to question (**45**, p. 173). In the second place, Holzinger (**45**, p. 254) maintains that the adjustment is grossly inaccurate when the obtained reliability is low. Thus, a reliability coefficient of 0.01 is increased to 0.75 when the σ is increased from 5 to 10. Finally, the evidence is against the assumption that the $\sigma_{meas.}$ remains constant throughout the range of learning ability. Spence (**110**) has shown that the $\sigma_{meas.}$ in the case of stylus maze scores does not remain constant.[4] Leeper (**64**) has shown that the $\sigma_{meas.}$ is a function of the amount of training and motivation in the learning of a maze by rats. In the case of the learning of nonsense syllables and words, there is suggestive, but not conclusive, evidence in Davis' (**26**) study that intra-individual variability in the learning of nonsense syllables and words increases with a decrease in the ability of the subject (an increase in the mean score). Also, it has been suggested (**13**, **14**) that intrinsic variability is very likely correlated with the level of ability of the subject in the task performed, and it is clear that such intrinsic variability is treated as an error of measurement in many of the methods used to determine reliability coefficients. Finally, the formula is invalidated if the errors of measurement in the 2 tests that yield the reliability coefficient are correlated (**45**, p. 254), and the correlation of errors of measurement is a common characteristic of the methods used to determine reliability coefficients for learning data.

In view of the questionable validity of the Kelley formula as a correction for the range of talent, Leeper (**64**) has emphasized the need for using comparable groups of subjects. He strongly recommends the use of the split-litter technique in comparisons of the reliability of methods used to study learning in rats. A comparable technique in the case of human subjects is the use of the same subjects throughout the experiment, or the use of subjects matched with respect to important factors such as age, sex, intelligence, amount of preliminary practice, etc. The many sources of error involved in the use of the same subjects in such comparative studies suggest the superiority of the matching technique. Perhaps the greatest need in the methodology of human learning is the definition of a standard group of subjects for use in all experimental studies of the reliability of different methods, materials, and measures. Inter-experiment comparisons would become possible if this were done.

(f) *Evaluation of the Specific Methods Used to Obtain Reliability Coefficients in Learning Experiments.* The general sources of

[4] Heron (**40**) failed to find any relation between variability and level of ability in stylus maze learning, except at the extremes.

variation in the reliability coefficients obtained for the methods and materials used in the study of learning have been indicated. We may now turn to the special methods that have been used, and attempt to evaluate their significance for use in developing a more precise methodology of experimental studies of human learning and memory.

(1) *Correlations Between Learning and Relearning Scores.* This method has been used rather infrequently both in studies of the reliability of animal learning scores (**39, 41, 64**) and of the reliability of materials used in studies of human learning (**40, 52, 90**).

In the case of rat learning, the correlations are very low. For example, Heron (**39**) found r's of 0.326 and 0.376 between the maze errors made by rats in series of trials that were separated by 175 days and 221 days, respectively. However, Leeper (**64**) has recently reported r's that range between 0.64 and 0.88 when rats are given 2 series of 6 trials on the same multiple-T maze with an interval of about 40 days between series. This suggests that the low correlations for the very long intervals may have resulted from differential changes in the learning abilities of the rats. In the case of human subjects the test-retest method frequently yields relatively high r's even though the intervals have not been exceptionally long. Thus, Hunter and Randolph (**52**) have reported r's of 0.49 and 0.58 for mazes and nonsense syllables when the intervals of rest were 160 days and 60 days, respectively; Hardin (as reported in Hunter, **50**) has obtained an r of 0.84 between learning and relearning scores separated by an interval of 84 days; Valentine and Meyer (**134**) have reported r's of 0.78 and 0.82, for men and women, respectively, between test and retest scores on the "lectometer" when the interval of rest was 30 days; and Peatman and Locke (**90**) have reported test-retest r's ranging between 0.35 and 0.50 for digit-span determinations separated by an interval of 60 days.

The methodological significance of these high r's is difficult to ascertain, but it may be suggested that they are partially attributable to the correlation between test errors, which should be high in view of the symbolic ability of the human subject, and partially attributable to a positive correlation between retentiveness and learning ability. Of some importance is the fact that retention was present in all the studies with human subjects despite the long intervals used. Furthermore, the lowest r's are those obtained with the digit-span test, and this probably represents the only type of learning test that can be correctly evaluated by the test-retest coefficient.

The usefulness of the learning-relearning correlation coefficient as an index in terms of which the reliability of different experimental methods may be compared obviously depends on the success with which retention can be eliminated. It has been suggested by Hunter and Randolph (**52**) that the learning-relearning r increases with the increase in the interval between maze and nonsense syllable tests, but this cannot be taken as evidence that an increase in the interval of rest necessarily results in a closer approximation of the r to the true reliability coefficient for the material used, *i.e.* that the r obtained when retention is present is necessarily too low. With early increases in the interval of rest, which are accompanied by decreases in the average retention of the group, the

r may increase merely because the relationship between learning ability and retention ability is revealed more clearly, *i.e.* the poorer learners may have returned to the state where they take just as long to relearn as they did to learn, whereas the better learners still retain something of what they originally learned and the spread between the good and poor learners is increased. Before the correlation may be considered as an index of the variable errors involved in measurement with a material, the means and $\sigma_{\text{dist.}}$'s obtained during the test and retest should be approximately equal; and this is the *sine qua non* whenever 2 methods or materials that may involve different degrees of retention are to be compared in terms of such test-retest coefficients. Even though these conditions are satisfied, it is necessary to consider the fact that the intrinsic variability of the abilities in question has been included as errors of measurement.

It should be noted that there are several sub-types of the learning-relearning method. In the first place, the interval of time may include no learning of similar materials under laboratory conditions (**41**, **52**, **64**), or other learning tests may be interpolated, as in Heron's study (**40**) of stylus maze learning and rational learning. The latter method introduces retroactive inhibition as a further aid to rapid forgetting of the primary material, and might for this reason be considered as a substitute for long periods of rest, but this conclusion is dangerous. The interpolation of a similar task affects recall scores on the primary task much more than saving scores (**76**, **79**), and the complete identification of the retroactive inhibition that results from specific interpolation of similar laboratory tasks and the oblivescence that occurs as a consequence of normal daily activities, is at present a legitimate hypothesis (**78**) but not a fact on which to base a methodology. The interpolation of similar learning between the learning and relearning tests may cause the relearning test to measure a third distinguishable ability, namely, the ability to overcome the interference effects.

The second variation in the learning-relearning method pertains to the method used to control the degree of learning in the first test. The subjects may either learn for a fixed number of trials or for a fixed amount of time (**39**, **41**, **52**, **64**), or they may learn to some criterion of mastery (**110**, **40**). Speculation as to the effects of these 2 methods on the learning-relearning r cannot be attempted, but it is clear that the use of a fixed number of trials has one advantage and several disadvantages. With a fixed number of trials for each subject the average degree of learning of the group must be lower than it would be if all subjects learned to the point of mastery—since the number of trials can be no greater than the number required for mastery by the most rapid learner—and this should hasten the forgetting of the material. On the other hand, the use of a fixed number of trials is not the typical procedure in learning experiments with human subjects, and the method cannot reveal the "true" reliability of the material or method as it is used by the experimentalist. It is fairly certain that the first few trials are not adequately representative of the learning abilities shown by records obtained for complete mastery. Furthermore, in the case of most of the materials used in the study of learning and memory with human subjects, the fastest learner requires not more than 4 or 5 trials for mastery, and this is too few trials for use in comparing the reliability of different materials.

(2) *The Correlation Between Measurements Obtained During the First and Second Halves of the Learning Period or During Various Groups of Trials.* The correlations between measures obtained in single trials, in groups of trials, or in the first and second halves of the total learning period, have been used as indices of the reliability of rat mazes (**64, 114, 122, 123**).

In the case of human learning, measures of this type have been used by Langdon (**62**) for evaluating the reliability of a simple motor learning test, by Nyswander (**87**) for comparing the reliability of stylus and finger mazes, and by Husband (**53**) for comparing the reliability of maze measures obtained with rats and human subjects. The significance of the method has been severely criticized by Leeper (**64**) and Spence (**110**), but the special case of the correlation between the total errors or time in the first and second halves of the learning period has been considered as one of the few valid methods by Tolman and Nyswander (**123**) and Nyswander (**87**).

It is apparent that the method, in any of its several forms that differ chiefly in the amount of data used, is merely a form of method (1) in which the interval between learning and "relearning" is made equal to the usual interval between trials or is filled with other trials on the same material. Therefore, the method has all the disadvantages of method (1) except that it eliminates the attenuating effect of quotidian variability. In the case of human subjects this effect is usually eliminated because the trials are all massed within a single learning period; in the case of rats this effect is usually eliminated (when several trials are grouped, or when the sum of the scores on the first half of the trials is correlated with the sum of the scores on the last half of the trials) by using the averages or sums of scores obtained on trials that are run on different days. However, the correlation of test errors is accentuated and a correlation of situation errors, such as are produced by distractions and other disturbances of the "average conditions of experimentation," is introduced. For these reasons, the correlations so obtained tend to overestimate the reliability of the method or material used.

But opposed to this overestimation is the very important underestimation of the reliability of the measurements obtained from the entire learning period as a consequence of the non-unitariness of the learning process. As previously indicated (p. 356), there is reason to believe that the different stages in the learning process represent basically different abilities, chiefly as a consequence of intra-serial interference effects and whatever characteristics of the material determine the occurrence of sudden insights. Further evidence is afforded by correlational analyses of the community of function represented by the measurements obtained in different trials or groups of trials during learning. The nearer the two sections of the learning period that yield the data, the higher the r obtained (**62, 64**). This may be taken either as evidence that there is a higher correlation between situation errors and test errors when the trials are adjacent, or that adjacent trials represent more nearly the same abilities.

In spite of this possibility that the scores obtained during the trials in the first and last halves of the learning period may represent different abilities, Tolman and Nyswander (123) and Nyswander (87) have maintained that the

method is adequate whenever "piecemeal" learning problems are employed. Likewise, Spence (110) has suggested that the exaggeration of the r due to the correlation between errors of measurement and the attenuation attributable to the measurement of different abilities may balance each other and give a fairly accurate estimate of the reliability of the maze that is *not* of the "piecemeal" type.

However, both proposed uses for such coefficients in reliability determinations suffer from the lack of an independent and quantitative criterion of "piecemealness." If a high r is obtained it may mean that the learning task is of the type that involves a progressive increase in the mastery of equal units of material and that the measurements are reliable, or it may mean that the task is not of the piecemeal type but that the correlation between errors of measurement is high; if the r is low, it may mean either that the maze is an unreliable example of the piecemeal type or that it represents the non-piecemeal type and has an indeterminate reliability.

There is another danger in using this r as an index of reliability. Without an independent criterion of the type of learning represented, the use of the r between test-halves in selecting materials and methods for experimentation, as Nyswander has done (87), implies that the investigator should use only those mazes, lists of verbal material, etc., that are learned in piecemeal fashion. This seems to be an illustration of Willoughby's point (139, p. 161) that the application of statistical criteria, which assume additive identical units, to problems of behavior sometimes ends in "the baby . . . being thrown out and the bath carefully studied." Such correlations may serve a useful function in the important task of identifying the type of learning being studied—it is undoubtedly important to know the extent to which the measurements obtained early in learning can be used to predict the measurements obtained late in learning—but they should never serve as indices to be used in selecting the most reliable methods and materials for the study of a particular type of learning.

Other difficulties encountered in the use of these methods for determining reliability support this conclusion. These difficulties are held in common with method (3), and may best be considered in the next section.

(3) *The Correlation Between the Sums of Scores on Odd and Even Trials During Learning.* The most extensively used index of the reliability of learning methods is the correlation between the sums of measurements (errors, correct responses, or times) on the odd and even trials during a specified group of trials on the same learning material (*e.g.* sums of measures obtained in trials $1+3+5+7 \ldots +19$ *versus* sums of measures obtained in trials $2+4+6+8 \ldots +20$). As variations of the method used to obtain this r there are: Spence's correlation between the sums of measures on the odd trials and the sums of measures on the even trials for whatever number of trials each subject requires to reach a criterion of mastery, *i.e.* a different number of trials contributes to the scores in the case of different subjects; and Anastasi's (2, 3, 4) correlation between the number of

errors or correct responses made during the first plus the third quarters and the second plus the fourth quarters of short work periods. This r is clearly another form of the test-retest coefficient, but one that is superior to the r's obtained by methods (1) or (2) since the data are fractionated in such a way that every part of the learning record contributes equally to both sets of records, and the measurements obtained are statistically comparable (see p. 359). Accordingly, the two series of measurements cannot represent different abilities or different stages in the learning process, and this source of systematic attenuation is eliminated. There have been several discussions of the exact meaning of the coefficient (123, 110, 64), and as a result it is most frequently labelled as an index of the *internal consistency* of the learning measurements.

The method seems to have been introduced by Heron (41) and has received explicit approval for use in learning studies by Tolman and Nyswander (123), Nyswander (87), Leeper (64), Anastasi (4), and Tryon (130). Tryon (130, p. 154) not only accepts the method, but has argued that "for such experimentally determined material as scores on successive trials in a maze, Spearman devised the method of collecting odd elements in one set, and even elements in the other set (the elements may be trials or blinds), correlating these two independent sets, and then applying the correction which eventually became known as 'Brown's formula' (*sic*). . . . This method, called the 'split-test' method, is the only exact method of determining the reliability coefficient of maze scores. All other methods, such as that of correlating the first half of the trials or blinds against the last half, or correlating one maze against another, do not satisfy the *definition* of the reliability coefficient, for the two sets of measurements used in the other methods are rarely comparable according to the criterion of comparability." This appeal to authority is, however, not valid, as Spence (110) has pointed out; nor is it quite accurate to identify this method with the well-known split-test method used in evaluating mental tests made up of a number of different items. Spearman defined the reliability coefficient as the correlation between the sums of scores made on odd and even comparable independent (non-identical) tests (see p. 361), whereas Tryon refers to the correlation of scores obtained from repetitions of the same material.

Nevertheless, the method has been widely applied by students of both rat and human learning. Thus, it has been used in comparing the reliability of mazes and other tests used in animal learning (39, 41, 55, 64, 114, 117, 123, 130), and in determining or comparing the reliability of stylus mazes (1, 87, 94, 105), finger mazes (6, 53, 87, 103), "lectometer" scores (134), scores obtained during verbal learning (2, 3, 4, 5), and the punchboard maze and Peterson Rational Learning Problem (38). It is of significance that the r's are exceptionally high when an appreciable portion of the total learning data is used in the calculations. Thirteen of 21 r's obtained from 6 or more trials with human subjects are above 0.90, and the lowest so far reported is 0.48. Some of these r's are so high that one is forced to suspect the existence of factors that tend to produce spuriously high coefficients. For example, Hall (38) has reported an

r of 0.990±.001 for odd-even correct recalls in the learning of a series of 22 nonsense syllables, and an r of 0.923±.011 for a sixteen-letter Peterson Rational Learning Problem. In the case of the Peterson Rational Learning Problem (eight-letter form) Garrison (36) has shown that the odd-even r is much higher than the r obtained from comparable forms (0.92 versus 0.70).

There is, of course, no statistical objection to repeating the same test over and over, provided each test is responded to by the subjects without reference, conscious or unconscious, to the earlier tests with the same material (139). The objection is that this procedure in the case of the learning experiment, where transfer of learning from trial to trial is the *sine qua non*, leads to a vitiation of the r as a measure of the variable error involved in the measurements. In all cases, the obtained r's are spuriously high as a consequence of the correlation between situation errors and the correlation between test errors. Furthermore, the method provides the proper mode of sampling for achieving a maximum correlation between these errors, particularly when practice is massed. In the case of the situation errors, any disturbing stimulus that has an effect that persists through more than one trial leads to a correlation between errors of measurement. Numerous examples of the correlation of test errors are available. In maze learning the position habits persist throughout several trials (64), and in verbal learning the chance meaningful associations affect the entire learning process. Furthermore, a frequent occurrence in the learning of nonsense syllables is the incorrect learning of a unit (*e.g. DOP* for *DOR*) as a consequence of a "chance" incorrect perception of the unit, and these incorrect responses may persist through several trials. The important point is that these sources of variable error are undoubtedly somewhat specific to particular kinds of mazes and particular kinds of verbal materials; yet their presence is not reflected in the odd-even reliability coefficients. One may wonder just what types of error the odd-even reliability coefficient measures, after these important types have been rather effectively eliminated from consideration.

Other difficulties are encountered in using this method or method (2). In the first place, these methods fractionate already scanty data. It is generally recognized that neither method yields a valid reliability coefficient if any subject in the group attains complete mastery of the material during the trials that are used to determine the coefficient (64, 75, 87, 110), because the r is then affected by the artificial failure to discriminate between some subjects. In memory studies this leads to a serious limitation in the number of trials available for the analysis of consistency, since many of the memory materials commonly used can be learned by some subjects in 4 or 5 trials.

This use of a fixed number of trials brings other problems in its wake. For example, in comparing 2 materials of different difficulty, should the investigator use the same number of trials in computing both reliability coefficients, in which case a greater part of the data obtained during the learning of the more difficult material is discarded, or should the investigator use the maximal number of trials obtainable in each case? Nyswander (87) chose the latter alternative in her study of stylus and finger mazes on the ground that the difference between the number of trials used was not great (16 and 10; 10 and 8), but then proceeded to correct all the correlation coefficients by using the Spearman-Brown formula. The solution of the problem may be correct, but the logic involves a

contradiction. Another closely related problem is the question of the accuracy with which the partial data predict the reliability of the scores obtained when the subjects learn to a criterion of mastery, as they most often do in experiments with human subjects. It may be suggested that a solution to both problems is to determine the correlation between measures obtained during a fixed number of trials and the measures obtained for complete learning in the case of each task to be compared, and then use in each case the number of trials that gives the prediction of the total scores with the same accuracy.

Another solution has been used by Spence (110). He had all his subjects satisfy a criterion of mastery in the stylus maze and then correlated the sums of the errors made on the odd trials and the sums of the errors made on the even trials during the entire learning period for each subject. Thus, the sums for subject *A* were obtained from, say, trials 1 through 10, and the sums for subject *B* were obtained from, say, trials 1 through 18. The *r*'s so obtained were higher than the *r*'s obtained by the usual method, a fact that Spence attributed to the use of more of the data. However, this method bears a close resemblance to Hunter's (41) use of Vincent curve values, and Leeper's (64, p. 166) objection that the latter method "spuriously raises the correlation by making the scores in any one tenth dependent on the total number of trials required by that subject to attain the norm of mastery" applies to Spence's method.

A second disadvantage of methods (2) and (3) is that the obtained reliability coefficients are based on not more than half of the available data. This has apparently not been considered as a disadvantage by most investigators, since they almost always correct for the halving of the data by using the Spearman-Brown Prophecy Formula. Only Leeper (64) and Spence (110) have questioned the justification for this bit of statistical manipulation. The formula is known to hold fairly well for non-learning tests, but its application to learning data of the sort involved in determining *r*'s by these methods may be seriously questioned until an empirical test has been made. The formula applies only in the case of strictly comparable tests that are psychologically independent of each other (25), and this is clearly not the case when inter-trial correlations are the crude *r*'s.

It must be concluded, chiefly on the basis of the known correlation between situation and test errors involved in these methods, that the *r*'s obtained by the correlation of scores on odd and even trials or for the first and second halves of learning lack the essentials of a good index of the variable error involved in measurements obtained with different experimental methods and materials. The question is not whether these *r*'s are overestimations or underestimations of the reliability of the measurements, but whether the overestimation and underestimation remain constant for different methods and materials. Such constancy seems improbable, and as a consequence a material that yields errors of measurement that are the least persistent and therefore least damaging may be judged less reliable than a material that breeds test errors and situation errors of the

type that affects performance throughout several trials or the entire learning period.

(4) *Correlations Between Scores on Comparable Samples of the Same Type of Learning Material or Task When the Samples Are Learned at Different Times.* The advantages and disadvantages of this method, and the definition of comparability, have been discussed previously (p. 358 ff.). The essentials of the specific method referred to here are (i) the use of strictly comparable materials that are neither so much alike that marked specific positive or negative transfer occurs nor so unlike as to measure different abilities; (ii) the learning of the 2 or more samples of the same material at different times, *i.e.* as discrete tasks.

It should be noted that the method need not involve the measurement of the intrinsic quotidian variability of the subject, since the 2 or more samples of some types of learning materials may be learned during the course of a single experimental session. Thus, this method has been used to determine the reliability of memory span methods (**90**) and to determine the reliability of measurements of immediate memory in studies of the intercorrelation of measures of memory and learning (**2, 3, 16, 35**). In these studies the subjects are usually presented with 4 groups of number or letter series that range from 4 to 10 units in length, and the correlated measures are the average span in the first and third groups and the average span in the second and fourth groups. Similarly, the reliability of measures of immediate memory for words, nonsense syllables, and other complex materials has been determined by repeating the test with comparable lists of words, etc., but method (4) has been used more often in such instances. The major difficulty encountered in the use of this method is that inter-serial positive and negative transfer occurs in the learning of the disparate lists or series, and the formally comparable lists become psychologically non-comparable. In the case of the memory span for words, Maslow (**74**) has demonstrated proactive inhibition when the lists are not separated by at least 40 seconds. It is, moreover, doubtful whether extended periods of rest between the presentation of such lists succeeds in eliminating this proactive transfer. Wyatt (**145**) has reported intrusions of members of the first list into the recall of a second list even though 6 weeks elapsed between lists.

A few attempts have been made to study the reliability of measures of the complete learning of complex materials by the use of comparable samples of the same materials. Thus, Heron (**40**) has studied the reliability of the stylus maze and the Peterson Rational Learning Problem by having the subjects learn 5 different stylus mazes and 2 Peterson Rational Learning Problems; Spence (**110**) has studied the reliability of the multiple-T stylus maze by having subjects learn 2 such mazes that were designed to be highly comparable; Stroud, Lehman, and McCue (**115**) have studied the reliability of lists of 6, 12, and 18 nonsense syllables by having each subject learn 2 lists of each length; and Garrison (**36**), Peterson and Telford (**97**), and Peterson and Lanier (**96**) have studied the reliability of the Peterson Rational Learning Problem by having subjects learn 2 comparable problems.

However, all of these studies are subject to the criticism that the 2 or more problems used were not strictly comparable. Independent determination, on control groups, of the comparability of the problems or materials was not made before the use of the materials in reliability determinations. Moreover, in the studies by Spence, Garrison, and Stroud, Lehman, and McCue the data show that the second tasks were learned more rapidly than the first. In Heron's study the second and third mazes were learned much more rapidly than the first, and the fourth and fifth were learned much less rapidly than the second and third. Perhaps the exceptionally low *r*'s obtained in this study may be explained as the effect of the intrusion of general habituation and specific positive and negative transfer, the negative transfer being cumulative and showing only after the general habituation had become relatively complete. The important point is that general habituation and specific transfer are known to occur, yet there have been no determinations of the reliability of these complex materials in which these factors have been eliminated or equalized.

The elimination or equalization of these factors as determinants of the 2 series of measurements that are to be correlated is the essential prerequisite for the interpretation of the resulting *r* as an index of variable error and for the use of such *r*'s in comparisons of different methods and materials. Since the *elimination* of these factors is accomplished not much more easily than the elimination of retention in the case of method (1), the most feasible procedure involves the *equalization* of the effects of habituation and specific transfer on the 2 series of measurements. Obviously, this equalization must be intra-subject, and must involve the learning of a number of comparable samples of the same material. The scores correlated may then be either (i) those obtained with 2 comparable samples of the material that have been learned late in the practice series, or (ii) the sums or averages of the scores obtained with 2 groups of several comparable samples of the material, the groups having been made up in such a way that practice and specific transfer are counterbalanced.

The significance of the correlations obtained under these conditions cannot be questioned. They represent an accurate estimate of the test errors and situation errors involved in the use of the materials in question, and if the materials lead to the measurement of approximately the same ability, the fact that the intrinsic quotidian variability of the subject is considered as an error of measurement does not vitiate comparisons of different experimental methods, materials, or measures. But there are several disadvantages. In the first place, the method that involves the equalization of practice and specific transfer yields reliability coefficients that are specific to materials learned by sophisticated subjects under conditions of near-maximal

positive or negative transfer of responses from past learning of similar material. The reliability of the scores obtained from naïve subjects cannot be validly determined except by the method that involves elimination of the retention of the first learning before administering the second test, and this is often impracticable. In the second place, the method is suited only for the use of the experimental methodologist; such extensive experimentation cannot be indulged in by the investigator who has a particular experimental problem to answer and needs a measure of reliability. Third, the method cannot be used with many learning materials or tasks because a number of comparable forms cannot be obtained. The importance of method (5) rests on these considerations.

(5) *Correlations Between Scores on Comparable Samples of the Same Type of Learning Material or Task When Both Samples Are Learned as Component Parts of a Single Material or Task.* In studying the reliability of the rat maze, Stone and Nyswander (**114**) introduced a variant of method (4) in which the r's were obtained by correlating the sums of errors made in the odd blinds with the sums of errors made in the even blinds, or the sums of the errors made in the last half of the maze with the sums of the errors made in the first half of the maze. The method is clearly analogous to the split-test method used in determining the reliability of non-learning tests, and the latter gives a crude r that represents the reliability of a task or material that is only half as long as the one actually used.

Up to now this method has been used only infrequently in the evaluation of the reliability of scores obtained with the complex materials used in studying human learning. Spence (**110**) has determined the r's for the errors in odd and even blinds in a stylus maze, and for the errors in the first and last half of the maze, and finds the former to be the higher of the two. Sackett (**103**) has reported an r of 0.67 between the errors made in the first and last half of a 24-cul finger maze. Stroud, Lehman, and McCue (**115**) have determined the r between the trials required to learn the first 6 and the last 6 nonsense syllables in a list of 12. They failed to find any difference between this r and the r obtained by correlating the trials required to learn 2 disparate lists of 6 nonsense syllables (r's$=0.61\pm.05$ and $0.65\pm.05$, respectively). They attributed this to the fact that there were pronounced individual variations in the order of learning; some subjects learned the syllables in progression from first to last and others learned the first few and last few units first, and learned the middle units last. Since this variation in the order of learning is probably largely a function of the set of the subject, the correlation between the first and last half of the units of a material is probably not a good index of the variable error in the measurements obtained from the entire material. The correlation between the scores on the odd and even units should be unaffected by these systematic differences between subjects, and is to be preferred for this reason.

In fact, this method has been the preferred one in all recent studies of immediate memory, and appears to be generally accepted as valid by students of the inter-relationships between measures of immediate memory and learning (2, 3, 16, 35).

The peculiar advantage of this method is that it yields an r which is comparable in some ways to the r obtained by the more laborious method (4) when practice and specific transfer are equalized, but without the disadvantage of restricting the determinations to measurements on sophisticated subjects. That is, the correlation between the sums of scores on odd units and on even units within a single material should be strictly comparable in so far as general habituation and specific interference and facilitation are concerned, and the intra-list transfer effects should be no more serious as a source of correlated test errors than are the inter-list transfer effects when many similar lists have been learned. Yet the split-test method is available for use with single samples of a material and may be used to study the effect of practice on the intra-test consistency.

However, this method is definitely inferior to method (4) in that it permits the correlation of situation errors, as has been noted by Leeper (64). Thus, a distracting stimulus may disturb the orientation of the subject and cause erroneous responses throughout an entire trial, and the fact that this same distracting stimulus may have a distracting effect of different duration when nonsense syllables or words are being memorized would not be revealed by correlations of the scores on odd and even units. Similarly, the differential effects of other accidental variations in the experimental situation, all of which are determinants of the obtained variability of measurements in experimental studies, could not be revealed. Perhaps this is not an insuperable obstacle to the use of this otherwise satisfactory method; the importance of the situation errors may have been overemphasized. The answer cannot be given until there have been studies in which 2 or more materials have been evaluated by both method (4) and method (5).

It is apparent that none of the several "reliability coefficients" computed for the measurements obtained in learning experiments is completely satisfactory as an index of the variable error involved in such measurements. The correlation between comparable forms of the same material seems to be the most promising method for the evaluation of different experimental methods and materials. But the use of this r requires the equalization or elimination of practice and specific transfer effects by elaborate experimental controls, and the comparisons of r's so obtained from different materials are valid indicators of the relative amount of accidental variation in the measurements only when the underlying ability may be assumed to be practically the same in the 2 cases; otherwise, differences in intrinsic variability play an important part in determining the variation of measurements.

The net result of these considerations is the generalization that

correlation coefficients obtained from learning data cannot be considered as reliability coefficients merely because they are correlation coefficients. The correlation coefficients may have great importance for the answering of certain questions regarding the nature of learning, but they cannot be considered as direct measures of the variable errors involved in individual measurements. Thus, the correlation between learning and relearning measurements, the correlation between the performance of subjects early in the learning period and late in the learning period, the correlation between the performance of subjects on the first half of a maze and on the last half of a maze are all important in their own right. But they cannot be considered as indices suited for use in evaluating the different experimental methods, materials, and measures used in the study of learning.

2. *The Absolute and Relative Variability of Measurements as Indices of the Reliability of the Methods Employed.* Davis (**26**), Sauer (**104**), McGeoch (**77**, **82**), and Stroud, Lehman, and McCue (**115**) have compared the reliability of materials and measures used in the study of verbal learning in terms of the absolute and relative variability of the measurements obtained. In these studies the variability of measurements obtained from the same subject (intra-individual variability) and the variability of measurements obtained from different subjects (inter-individual variability) are used as commensurate indices. Thus, Davis (**26**) had each of 6 subjects learn 20 lists of 12 nonsense syllables and 20 lists of 12 three-letter words to a criterion of 1 errorless trial. The subjects learned 1 list each day and were given 2 days of preliminary practice. The relative reliability of the lists of words and nonsense syllables was then determined by comparing the relative variability (V) of the trials required to learn the words and the relative variability of the trials required to learn the nonsense syllables in the case of each subject. On the other hand, Sauer (**104**) compared the reliability of nonsense syllables and three-letter words in terms of the relative variability of the number of trials required by 20 subjects to learn a list of 24 nonsense syllables and a list of 24 words. Each subject learned 5 lists of words and 5 lists of nonsense syllables, so that the individual-to-individual variability in the difficulty of lists of words and lists of nonsense syllables could be determined at different stages of practice. The other studies mentioned were similar to Sauer's study except that Stroud, Lehman, and McCue (**115**) and McGeoch (**77**) used a counterbalanced practice order for their different conditions,

that is, different subjects learned the same lists after different amounts of practice in learning.

The logic of the use of the $\sigma_{dist.}$ in comparative studies of the reliability of experimental methods is that, *ceteris paribus,* the reliability of an obtained mean or difference between means varies inversely as the $\sigma_{dist.}$. If, therefore, one experimental method yields less variable individual measurements than another, the numerator in the equation for the σ_{mean} ($\sigma_{dist.}/\sqrt{N}$) is smaller, and the error involved in the estimate of the "true" mean is decreased. Stated in another way, the experimental method that yields the least variable individual measurements must, if our statistical formulae are correct, yield means from successive samples of N observations that are most nearly the same. Or the argument for the use of a measure of obtained variability may rest on an analysis of the causes of variance rather than on the assumed relationship between the σ_{mean} and the $\sigma_{dist.}$. The $\sigma_{dist.}$ obtained in any experimental study is a resultant of the variable errors that may be attributed to the inconstancy of each aspect of the experimental situation. Thus, $\sigma_{dist.}=\sqrt{\sigma^2_a+\sigma^2_b+\sigma^2_c \ldots +\sigma^2_x}$, where σ^2_a, σ^2_b, $\sigma^2_c \ldots \sigma^2_x$ are the variances attributable solely to the inconstancy of factors $a, b, c, \ldots x$. If, therefore, all factors in the experimental situation are the same except one, which is the nature of the material learned, and the total variability is greater with material A than with material B, it may be concluded that material B contributes less to the total variability than material A, and that material B is more constant in difficulty than material A. Similarly, if 2 complex and unanalyzed experimental situations, A and B, yield different $\sigma_{dist.}$'s, the situation that gives the smaller $\sigma_{dist.}$ may be said to be more reliable either because the factors involved in it are more constant or because fewer variable factors are present.

A major difficulty in the use of the $\sigma_{dist.}$ as an index of reliability occurs when there is a shift in the unit of measurement. Obviously, the relative reliability of time and error scores cannot be determined by comparing the $\sigma_{dist.}$ of the time scores and the $\sigma_{dist.}$ of the error scores. Likewise, it is generally assumed that the σ's of 2 distributions that center about different mean values cannot be directly compared, even though the unit of measurement used is ostensibly the same, *i.e.* trials, time, or errors. Thus, if one experimental method or material yields a mean of 10 trials and $\sigma_{dist.}$ of 2, and another method or material yields a mean of 15 trials and $\sigma_{dist.}$ of 3, there arises the question whether the "actual" reliability, *i.e.* freedom from variable error, of the second is less than that of the first.

Without exception the investigators (**26, 77, 82, 104, 115**) who have attempted to compare the reliability of different memory materials and methods in terms of the variability of obtained learning scores have assumed that the proper index of reliability is not the $\sigma_{dist.}$ but the ratio of the $\sigma_{dist.}$ to the mean, as expressed in Pearson's Coefficient of Relative Variability, V. These investigators have, however, failed to present detailed justifications for assuming that V is a valid measure of the reliability of a method. For example, Davis (**26**, p. 225) says merely: "In scientific work . . . we are primarily concerned with the reliability of the difference between two means and with the possibility of predicting whether the same kind of difference will be obtained if the experiment is repeated. For this purpose, the relative reliability of the mean, that is, the ratio of the mean to the standard deviation, is the really significant measure." That this relationship is axiomatic is questionable. At least, the rather extensive controversial literature on the problem of absolute *versus* relative variability in studies of the effect of practice on individual differences (**5, 29, 51**) suggests that the use of V as an index of reliability needs critical examination.

The assumptions involved in the use of V as an index of the reliability of a method are apparent when Davis' statement is elaborated. Let it be supposed that under condition A subjects require a mean of 15 trials for mastery of a ten-unit list of nonsense syllables, with a $\sigma_{dist.}$ of 3, and that the same subjects require a mean of 10 trials to master a ten-unit list of words, with a $\sigma_{dist.}$ of 2. The adherents of V as a measure of reliability must consider the 2 materials to be equally reliable, because the V is 20.0 in each case. In considering them equally reliable, it is implied that the 2 materials must give equally reliable differences between the experimental condition A and another experimental condition, B. However, the $\sigma_{dist.}$'s, and not the V's, are employed in computing the reliability of differences between means. *Therefore, the 2 materials could give equally reliable experimental differences only in case the obtained mean difference between conditions A and B increased in direct proportion to the increase in the* $\sigma_{dist.}$. In the example cited above, if conditions A and B gave means of 10 and 15 ($\sigma_{dist.}=2$ and 3) when nonsense syllables were used, then conditions A and B must give, according to the adherents of V, means of 5 and 7.5 ($\sigma_{dist.}=1$ and 1.5) when the lists of words are used. N is, of course, considered constant. If $N=25$ in each case, the obtained differences would in both cases be equal to 6.94 $\sigma_{diff.}$, and this is the fact which the equality of the V's is supposed to indicate.

Two distinct assumptions are involved in this use of V, and both have been criticized by students of the relationship between practice and individual differences. (1) The first assumption is that there is a perfect positive correlation between the magnitude of a mean, the

magnitude of the σ of the distribution centered at that mean, and the magnitude of the change in that mean that occurs when the experimental conditions are altered. In short, a change in the value of a mean score is thought to involve a necessary change in the units of measurement which is reflected in the magnitude of the $\sigma_{dist.}$ and in the magnitude of the effect of an alteration in the experimental conditions. The assumption that *some* change in the $\sigma_{dist.}$ must occur is generally considered valid, and the evidence is preponderantly confirmatory. As pointed out by numerous writers, but especially Peterson (93, 95), an increase (or decrease) in the mean score is almost invariably accompanied by an increase (or decrease) in the $\sigma_{dist.}$ of the scores. The classic example is the learning curve plotted in terms of time per unit of work and in terms of work per unit of time. In the first case, the mean score and the $\sigma_{dist.}$ decrease with practice, and in the second case, the mean score and the $\sigma_{dist.}$ increase with practice.

Although this is an excellent illustration of the impossibility of using absolute variability as an index of the relative reliability of memory methods that yield different mean scores, Anastasi (5) questions the inevitability of the change in the $\sigma_{dist.}$ when the mean changes. Her argument is that the correlation between the means and $\sigma_{dist.}$'s obtained in successive practice periods should be +1.00, if there are concomitant changes in both measures that are attributable to a change in the unit of measurement. But she obtained r's in the learning of cancellation, hidden words, symbol-digit substitution, and vocabulary that were only 0.79, 0.95, 0.70, and 0.74, respectively. In addition, an increase in the mean test score was found to be accompanied at the end and beginning of practice by no increase or by a decrease in the $\sigma_{dist.}$. Anastasi cites these facts as evidence against the assumption made by Peterson and others. But this does not seem to be a necessary conclusion. Anastasi's data merely show that psychological factors may influence the mean and $\sigma_{dist.}$ independently. This is not damaging to the assumption as it has been used by students of methodology. The problem of determining the reliability of different methods results from the assumption that the $\sigma_{dist.}$ was not wholly determined by the size of the mean. If the use of V in this connection has any justification, it must be assumed that the size of the mean *and* "psychological" factors determine the obtained $\sigma_{dist.}$ and that the r between the mean and the $\sigma_{dist.}$ is +1.00 only when the effect of "psychological" factors is partialed out.

It is the second of the two assumptions made by those who use V

as an index of reliability that is definitely unwarranted and contrary to the facts. In the first assumption, the assertion was merely that the size of the $\sigma_{dist.}$ and the size of the mean vary together; the second assumption is that the ratio between the *obtained* $\sigma_{dist.}$ and *obtained* mean, and the ratio between the obtained mean and a change in the mean produced by a change in the experimental conditions, is a constant. To return to our example, if the mean of the trials for learning a ten-unit list of words is 5, and an experimental change produces an increase of 2.5 trials, it is then assumed that the same experimental change would increase the trials for learning nonsense syllables from 10 to 15. Similarly, when the mean trials for learning the list of words is 5, and this is accompanied by a $\sigma_{dist.}$ of 1, it is assumed that the list of words would have yielded a $\sigma_{dist.}$ of 2, if, somehow, the mean trials for learning words had been 10 rather than 5. The fundamental error in these extrapolations has been clearly stated by Thurstone (**118**) and reviewed by Anastasi (**5**). It is that such extrapolations assume that the mean has been measured from an absolute zero, and we have no assurance that absolute zero-points are used in measuring performance in memory and learning experiments. Assuming that the ratio between $\sigma_{dist.}$ and the mean as measured from absolute zero is a constant, Anastasi has shown that the relation between the *V*'s computed for 2 experimental conditions may depend entirely on whether an arbitrary or absolute zero-point has been used in determining the scores.

Although it is doubtful whether the constancy of the ratio between the $\sigma_{dist.}$ and the mean can be determined empirically, since the mean under a particular condition cannot be altered without altering some part of the experimental situation, it is possible to check the assumed constancy of the relationship between the size of the mean obtained and the magnitude of the effect produced by an experimental variation in conditions.

Pertinent data are presented in the study in which Davis (**26**) makes the assumption that a constant relationship exists. As previously indicated, each subject learned 20 lists of nonsense syllables and 20 lists of words, and means and $\sigma_{dist.}$'s were computed separately for each subject. If such subject's ability in learning nonsense syllables and words is considered as a fixed experimental deviation from the mean of the group of 6 subjects, an analysis of Davis' Table I (**26**, p. 225) reveals the group means for the learning of words and syllables to be 3.74 and 9.80, a ratio of 1.00 to 2.62, whereas the *M.V.*'s of individual means from the group means were 0.83 and 0.68, respectively, a ratio of 1.00 to 0.83. The two ratios should have been the same, if the assumption regarding the validity of *V* as an index of reliability is sound. In conclusion,

it must be stated that the use of V as an index of the reliability of experimental differences which would be obtained if a particular method were used involves extrapolations which are theoretically and empirically unsound. Equal V's do not necessarily indicate that equally reliable experimental differences will be obtained.

The inadequacy of V leaves only 2 alternative ways in which direct measures of variability, either inter- or intra-individual, may be used for comparing the reliability of methods which yield different mean scores. Both methods substitute direct experimental control and measurement for specious statistical short-cuts. The first method involves the experimental manipulation of certain factors in the situation until the 2 experimental methods yield the same mean scores, and the $\sigma_{dist.}$'s may be directly compared. Thus, the discovery that 10 words are learned in 10 trials and that 10 nonsense syllables are learned in 15 trials need not stifle the comparison of the reliability of nonsense syllables and words. It is as legitimate to require that the 2 materials be compared when they require the same number of trials for mastery as to require that the same number of units of each material be used. In short, equality of mean scores may be set as a criterion which is to be met before comparisons of variability are made. Obviously, this technique is inapplicable to the comparison of the reliability of different measures of learning and memory, such as trials, time, or errors.

The second method is to express the measure of absolute variability as a per cent of the obtained mean differences between experimental conditions or subjects. This method has the advantage that it results in an expression of the particular relationship in which the experimentalist is interested. It is fundamentally the same as the correction attempted by using V, but with the important difference that the validity of this method does not depend on either the assumption that the difference between means increases as the means increase, or on the assumption that measurements have been made from an absolute zero. The differences between the means for different experimental conditions or subjects, most conveniently expressed in the σ of the distribution of the means obtained with different conditions or subjects, are not affected by the use of an arbitrary zero. The values are differences between positive amounts rather than deviations from the zero of a scale. Furthermore, the validity of the proposed ratio does not depend on an assumed constant ratio between the size of the mean and the extent to which the mean is changed by an alteration in the experimental conditions, because the changes

actually produced by different conditions are measured. In practice, this ratio of the $\sigma_{dist.}$ to the σ of the distribution of means obtained under different conditions may take 2 forms. (a) The $\sigma_{dist.}$ measures the variability of a single subject in successive performances of the same task, and the σ of the distribution of the means obtained under different "conditions" is the σ of the distribution of the means obtained from different subjects. (b) The $\sigma_{dist.}$ measures the inter-individual variability under a single condition, and the denominator of the ratio is the σ of the distribution of means obtained under different experimental conditions with the same group of subjects or comparable groups of subjects. The first will be immediately recognized as essentially the same measure of reliability as that given by the reliability coefficient. The second measure has not been applied in any comparisons of techniques to date, but the first has been used by Davis (**26**).

Sources of Error in the Use of Measures of Variability as Indices of Reliability. The justification for the use of these indices to compare different methods and materials in terms of the variability of measures obtained with them depends on the adequate control of sources of error similar to those discussed in connection with the reliability coefficient. Three of these vitiating factors deserve special comment, since they have been present in the studies that have used the indices in question.

(1) *Intrinsic Variability.* In all the studies that have used measures of absolute or relative variability, quotidian variability has been one component of the total variance. The intrusion of quotidian variability, which is attributable partly to intrinsic variability and partly to a lack of constancy in the experimental conditions, may, as Anastasi (**5**) has suggested, lead to the interpretation of intrinsic variability as unreliability. However, as previously noted (p. 354 ff.), it is improbable that the true intrinsic variabilities of the traits underlying the learning of, say, nonsense syllables and words, are greatly different. Furthermore, the experiments of Davis (**26**) *et al.* have attempted to simulate the actual conditions of experimentation on memory, and have *prima facie* validity for the purpose stated.

(2) *Artificial Restriction of Range.* The $\sigma_{dist.}$ measures the reliability of a method or material only when it can be assumed or demonstrated that the method or material is adequate for measuring intra-individual differences and inter-individual differences throughout the entire range of each. If for any reason the subjects above or below a certain point in the distribution of true abilities receive the same score when a certain method is used, the $\sigma_{dist.}$ is smaller than it would have been if the method had been adequate for measuring the entire range of true abilities. Similarly, the measures of intra-individual differences are attenuated when the range of differences is artificially curtailed. For example, in McGeoch's (**77**) study an attempt was made to determine the

relative reliability of three-letter words and nonsense syllables of different degrees of meaningfulness. The subjects learned lists of 10 units of each material for 2 minutes and attempted a recall immediately thereafter. The mean recall for the three-letter words was 9.00, with a σ of 1.12, and the mean recall for the nonsense syllables was 7.35, with a σ of 1.96. Although the actual distributions of the scores are not given, it is clear that the distribution of the scores obtained from the learning of the three-letter words must have shown a marked concentration of cases at the upper limit, *i.e.* 10 units, and that the distribution of scores obtained from the learning of the nonsense syllables must have been curtailed to a lesser degree. Accordingly, the obtained σ's (or V's) cannot be accepted as indices of the reliability of the materials employed. If they were to be accepted, one must necessarily conclude that the least difficult learning material is the most reliable material. The same artifact may enter into any studies of reliability in which the materials are so difficult that few subjects are able to master any of the material in the time allotted, those in which the criteria of mastery are not sufficiently severe to eliminate zero scores, or those in which the criteria of mastery are so severe that some subjects never achieve mastery. The point is that whenever a measure of variability is used to compare the reliability of experimental methods, the comparison must be accompanied by evidence that the range of scores obtained by the different methods has not been artificially restricted. The most acceptable evidence is perhaps the obtained distributions, since it is conceivable that a distribution satisfy a test for normality and yet show a sensible restriction of range, or that it fail to satisfy such a test and yet show no sensible restriction of range.

(3) *The Validity of the Assumed Relation Between* $\sigma_{dist.}$ *and* σ_{mean}. A final consideration in the use of measures of variability as indices of the reliability of experimental methods is the extent to which the investigator may justifiably assume that the true variability of the means obtained from samples of N observations may be accurately estimated from the known $\sigma_{dist.}$ and N of the single sample on the assumption that the Lexian ratio equals 1. It is apparent that the use of the indices outlined in the preceding pages assumes (a) that the σ_{mean} may always be accurately estimated when the $\sigma_{dist.}$ is known, and (b) that the estimate is equally accurate in the case of the 2 experimental methods to be compared. Neither assumption is valid unless the single observations included in the sample have been drawn at random from a homogeneous parent population. As previously indicated (p. 342 ff.), overestimation of the σ_{mean}, or underestimation of the reliability of the obtained mean, occurs whenever the observations *included* in the single sample represent a heterogeneous parent population (such that some of the observations come from sub-group 1, some from sub-group 2, some from sub-group 3, etc., each subgroup having a different mean), and the degree of overestimation of the σ_{mean} is a function of the amount of difference between the means of the subgroups represented. If, therefore, the observations obtained in comparative studies of experimental methods or materials are from heterogeneous populations, the usual formula for estimating the σ_{mean} is not accurate, and a comparison of the reliability of different methods may be completely vitiated if the parent populations from which the samples are drawn are not *equally* heterogeneous.

UNIVERSITY OF MICHIGAN LIBRARIES

Heterogeneity of the populations from which the series of observations are drawn may be effectively eliminated or held at a constant value in so far as many factors are concerned. Thus, sex and age differences in memory ability need present no particular problem. But progressive changes in the performance of subjects, such as produced by practice and fatigue, may vitiate the comparison of the reliability of learning materials or methods if the investigator fails to reduce such changes to a minimum before beginning his observations. It is customary to assume that methods of counterbalancing experimental conditions effectively eliminate practice effects as a source of experimental error, even though the effects of practice are still noticeable. But this assumption is valid only when the mean performance under conditions *A*, *B*, etc. is under consideration; counterbalancing destroys the significance of measures of variability. If practice effects are operating to increase the speed of learning from test to test, it is apparent that the $\sigma_{\text{dist.}}$ does not measure the deviations of scores due to chance factors alone. In effect, the learning scores of a subject on successive days represent samples from a heterogeneous population; on day 1 the score represents sub-group 1 which has a mean of, say, 20; on day 2 the score represents sub-group 2 which has a mean of, say, 15; on day 3 the score represents sub-group 3 which has a mean of, say, 12; etc. The series of observations obtained under such conditions does not satisfy the assumptions involved in the use of the formula for the σ_{mean}; the $\sigma_{\text{dist.}}$ obtained from the series underestimates the reliability of the obtained mean by an unknown and indeterminate amount; and there can be no justification whatsoever for the comparison of the reliability of 2 materials or methods unless it can be shown that the practice effects are the same for the 2 materials or methods. The practice effect from list to list is known to vary greatly as a function of the material learned, and probably varies as a function of the method of presentation, etc. (see p. 359 ff.). Therefore, the only effective method for assuring the meaningfulness of comparisons of the reliability of different methods and materials in terms of an index of variability involves at least an approximate elimination of practice effects before the measurements are made. This is particularly true in studies of intra-individual variability. Some of the comparisons made by Davis (**26**), McGeoch (**77**), and Stroud, Lehman, and McCue (**115**) are subject to criticism on this point.

3. *The Variability of Means from Successive Samples Obtained Under the Same Experimental Conditions as an Index of the Reliability of the Methods, Materials, and Measures Used.* Maurer and Carr (**75**) have recently determined the relative reliability of different measures of learning and criteria of mastery in the maze learning of rats in terms of a third index of reliability [5], namely, the ratio of the obtained difference between the means from samples obtained under the same experimental conditions to the *P.E.* of the difference

[5] Heron (**40**) used a somewhat similar method of analysis in his study of the reliability of stylus maze measures with human subjects, but did not base his conclusions on the ratio: difference/$P.E._{\text{diff.}}$.

as computed by the formula $P.E._{diff.} = \sqrt{P.E.^2_{mean\ x} + P.E.^2_{mean\ y}}$, where $P.E._{mean} = .6745\sigma_{dist.}/\sqrt{N}$. After having controlled the experimental conditions with great care, 3 groups of rats were run on a nine-cul Carr maze until they reached a criterion of 8 errorless runs in 10. The records were then analyzed to determine the mean time, trials, and errors required by each group to reach each of 3 criteria of mastery—2 errorless trials in 3, 4 errorless trials in 5, and 8 errorless trials in 10. The critical ratios were then computed using the differences between the means for trials, time, and errors for each of the 3 groups of rats and for each of the 3 criteria of mastery. It was found that 7 of 9 differences between the means for time scores were greater than 3 times their respective *P.E.*'s; 3 of 9 differences between the means for error scores were greater than 4 times their respective *P.E.*'s, and 4 differences were between 2 and 3 times their *P.E.*'s; and that only 2 of the differences between the means of trial scores were between 3 and 4 times the *P.E.*'s, and only 1 difference was between 2 and 3 times its *P.E.* They concluded that the trial scores were the most reliable and that the error scores were the least reliable.

The basic conception underlying this method for determining the reliability of experimental measurements is undoubtedly valid. It is, in short, a method for determining directly the facts other investigators have attempted to predict by means of the $\sigma_{dist.}$, V, or reliability coefficient, namely, the amount of chance variation to be expected between means of measurements obtained with a particular method, material, or measure. However, one may question whether the critical ratio is the proper or sufficient index for comparing different methods, etc., in terms of the amount of variable error permitted when they are used.

In the first place, Maurer and Carr (75) *have actually compared the accuracy of their statistical estimates of error based on the assumption of random sampling, rather than the relative amounts of variable error involved in time, error, and trial measurements.* That is, they have determined that the Lexian ratio (see p. 341) for error measurements is greater than 1 and that the Lexian ratio for trial measurements is less than 1. The similarity between this experiment and Woodrow's (143) and Culler's (24) is striking. If time, error, and trial measurements had been selected at random from homogeneous populations, the critical ratios would not have been significantly different—in the long run—for these 3 measures, even though the actual amounts of variable error involved in the measures differed

UNIVERSITY OF MICHIGAN LIBRARIES

greatly. Thus, the intra-sample variability could be much greater in the case of trial measurements than in the case of error measurements, yet the obtained critical ratios between the means of successive samples would be equal in average size so long as the Lexian ratio was 1 in both cases.

This restriction of the meaning of the results obtained by Maurer and Carr is not a criticism. Studies of the accuracy with which the usual statistical formulae predict the actual variation between means from successive samples are extremely important for the experimentalist, and the *raison d'être* of such studies has been stated formally as our first sub-criterion of reliability (see p. 339). But there remains the problem of determining the actual reliability of the means of measurements obtained with different methods, materials, and measures, *i.e.* the actual variation between means from successive samples.

In order to determine the relative variability of the means obtained from successive samples when 2 different materials, etc., are used, the index must be independent of the units of measurement in terms of which the obtained σ's are expressed. To accomplish this by using the ratios of the obtained standard deviations of the sample means to the respective means of all observations is not satisfactory, because it assumes that the measurements have been taken from the absolute zero. Likewise, the average ratio of the inter-sample differences to the estimated $\sigma_{diff.}$ (or $P.E._{diff.}$) is not satisfactory because it assumes that the samples have been obtained from equally homogeneous parent distributions in the 2 instances being compared. The solution must be similar to the one stated in the preceding section (p. 381), namely, the standard deviation of the means of successive samples must be expressed in terms of the average difference between means obtained under experimental condition X and experimental condition Y. When this is done, the ratios obtained are independent of the units of measurement, independent of the assumptions of simple sampling, and independent of the assumption of an absolute zero for the obtained measurement. If, therefore, 2 different materials (methods, measures) are used under Conditions X and Y, and the entire experiment is repeated a number of times, the actual frequency of differences greater or less than that to be expected on the basis of the obtained variation between means obtained under like conditions is the final and unquestionably valid index of the relative reliability of the 2 materials. The material that yields reliable differences most often is necessarily the most sensitive.

The results of this method must stand as the final criterion for evaluating different experimental methods in terms of the variable error in the measurements obtained when they are used, and as the criterion in terms of which the other short-cut methods for evaluating variable errors must be validated.

III. THE CRITERION OF CONFORMITY

The extraordinary complexity of the problems involved in the determination of the validity and reliability of the many aspects of the methodology of human learning and memory undoubtedly explains the present failure of investigators to achieve a standardization of their procedures. Moreover, the empirical evaluation of the different methods, materials, and measures will probably proceed at a slow pace as a consequence of the unrelieved complexity of those problems of statistical analysis. But the need for standardization of techniques is pressing, and should not be frustrated by the slow progress of an experimental approach to methodology.

In view of this, any attempt to summarize and evaluate the methods, materials, and measures in use at the present time in studies of human learning and memory must appeal frequently to a third criterion, namely, *the frequency with which an experimental method, material, or measure has been used in other experiments in the past.* This criterion should not, however, be interpreted as another quantitative criterion. It is important to discover that a three-second presentation interval for nonsense syllables has been used by 51 investigators and that a two-second presentation interval has been used by 35 investigators, but it is of greater importance to know that the two-second interval has been used in a number of studies in which basic variables have been investigated. That is, the conditions used in an experiment, such as the one performed by Luh (**68**), must be weighted heavily in such evaluations because they were the basis for generalizations that are involved in almost every study of memory.

BIBLIOGRAPHY

1. Allison, L. W., Difficulty as a Factor in the Standardization of a Maze. *J. Gener. Psychol.*, 1931, **5**, 514–518.
2. Anastasi, A., A Group Factor in Immediate Memory. *Arch. of Psychol.*, 1930, **18**, No. 120. Pp. 61.
3. Anastasi, A., Further Studies of the Memory Factor. *Arch. of Psychol.*, 1932–1933, **22**, No. 142. Pp. 60.
4. Anastasi, A., The Influence of Practice on Test Reliability. *J. Educ. Psychol.*, 1934, **25**, 321–335.

UNIVERSITY OF MICHIGAN LIBRARIES

5. ANASTASI, A., Practice and Variability. *Psychol. Monog.*, 1934, **45**, No. 204. Pp. 55.
6. BARKER, R. G., A Temporal Finger Maze. *Amer. J. Psychol.*, 1931, **43**, 634–636.
7. BARR, A. S., A Study of the Amount of Agreement Found in the Results of Four Experimenters Employing the Same Experimental Technique in a Study of the Effects of Visual and Auditory Stimulation in Learning. *J. Educ. Res.*, 1932, **26**, 35–45.
8. BARTLETT, F. C., Experimental Methods in Psychology. *J. Gener. Psychol.*, 1930, **4**, 49–66.
9. BOOK, W. F., Psychology of Skill. *U. Mont. Stud. in Psychol.*, 1908, Bull. No. 53, Psychol. Ser. I.
10. BORING, E. G., Mathematical *vs.* Scientific Significance. *Psychol. Bull.*, 1919, **16**, 335–338.
11. BORING, E. G., Scientific Induction and Statistics. *Amer. J. Psychol.*, 1926, **37**, 303–307.
12. BROWN, W., Some Experimental Results in the Correlation of Mental Abilities. *Brit. J. Psychol.*, 1909–1910, **3**, 296–322.
13. BROWN, W., The Effects of 'Observational Errors' and Other Factors upon Correlation Coefficients in Psychology. *Brit. J. Psychol.*, 1913–1914, **6**, 223–238.
14. BROWN, W., and THOMSON, G. H., *The Essentials of Mental Measurement.* Cambridge: Cambridge University Press, 1925. Pp. 224.
15. BRUCE, R. W., Conditions of Transfer of Training. *J. Exper. Psychol.*, 1933, **16**, 343–361.
16. BRYAN, A. I., Organization of Memory in Young Children. *Arch. of Psychol.*, 1933–1934, **24**, No. 162. Pp. 56.
17. BRYAN, W. L., and HARTER, N., Studies in the Physiology and Psychology of the Telegraphic Language. *Psychol. Rev.*, 1897, **4**, 27–53; 1899, **6**, 345–375.
18. CARR, H. A., The Reliability of the Maze Experiment. *J. Comp. Psychol.*, 1926, **6**, 85–93.
19. CARR, H. A., Teaching and Learning. *J. Genet. Psychol.*, 1930, **37**, 189–218.
20. CARR, H. A., The Quest for Constants. *Psychol. Rev.*, 1933, **40**, 514–532.
21. COMMINS, W. D., MCNEMAR, Q., and STONE, C. P., Intercorrelations of Measures of Ability in the Rat. *J. Comp. Psychol.*, 1932, **14**, 225–235.
22. COREY, S. M., A Summary of Certain Factors in Current Investigations on Learning and Memory. *Amer. J. Psychol.*, 1932, **44**, 190–192.
23. CRAFTS, L. W., Transfer as Related to Number of Common Elements. *J. Gener. Psychol.*, 1935, **13**, 147–158.
24. CULLER, E., Studies in Psychometric Theory: XIV. On the Probable Error of the Limen (Method of Constant Stimuli). *J. Exper. Psychol.*, 1927, **10**, 463–477.
25. CURETON, E. E., Errors of Measurement and Correlation. *Arch. of Psychol.*, 1930–1931, **19**, No. 125. Pp. 63.
26. DAVIS, F. C., The Relative Reliability of Words and Nonsense Syllables as Learning Material. *J. Exper. Psychol.*, 1930, **13**, 221–234.

27. DUNLAP, J. W., Comparable Tests and Reliability. *J. Educ. Psychol.*, 1933, **24**, 442–460.
28. EBBINGHAUS, H., *Memory.* (Trans. by H. A. Ruger and C. E. Bussenius.) New York: Bureau of Publications, Teachers College, Columbia University, 1913.
29. EWERT, P. H., The Effect of Practice on Individual Differences When Studied with Measurements Weighted for Difficulty. *J. Gener. Psychol.*, 1934, **10**, 249–285.
30. FISHER, R. A., *Statistical Methods for Research Workers.* (Fifth Ed.) Edinburgh: Oliver and Boyd, 1934. Pp. 319.
31. FISHER, R. A., *The Design of Experiments.* Edinburgh: Oliver and Boyd, 1935. Pp. 250.
32. FOUCAULT, M., Les associations locales et la loi de fixation des images. *Ann. psychol.*, 1929, **30**, 25–39.
33. GARRETT, H. E., The Relation of Tests of Memory and Learning to Each Other and to General Intelligence in a Highly Selected Adult Group. *J. Educ. Psychol.*, 1928, **19**, 601–613.
34. GARRETT, H. E., The Two-Factor Theory and Its Criticism. *Psychol. Rev.*, 1935, **42**, 293–301.
35. GARRETT, H. E., BRYAN, A. I., and PERL, R. E., The Age Factor in Mental Organization. *Arch. of Psychol.*, 1934–1935, **26**, No. 176. Pp. 31.
36. GARRISON, K. C., An Analytic Study of Rational Learning. *George Peabody Coll. for Teachers, Contr. to Educ.*, 1928, No. 44. Pp. 52.
37. GLAZE, J. A., The Association Value of Non-sense Syllables. *J. Genet. Psychol.*, 1928, **35**, 255–269.
38. HALL, C. S., Intercorrelations of Measures of Human Learning. *Psychol. Rev.*, 1936, **43**, 179–196.
39. HERON, W. T., The Test-Retest Reliability of Rat Learning Scores from the Multiple-T Maze. *J. Genet. Psychol.*, 1930, **38**, 101–113.
40. HERON, W. T., Individual Differences in Ability *Versus* Chance in the Learning of the Stylus Maze. *Comp. Psychol. Monog.*, 1924, **2**, No. 8. Pp. 60.
41. HERON, W. T., and HUNTER, W. S., Studies of the Reliability of the Problem Box and the Maze with Human and Animal Subjects. *Comp. Psychol. Monog.*, 1922, **1**, No. 1. Pp. 56.
42. HICKS, V. C., and CARR, H. A., Human Reactions in a Maze. *J. Anim. Behav.*, 1912, **2**, 98–125.
43. HILGARD, E. R., The Saving Score as a Measure of Retention. *Amer. J. Psychol.*, 1934, **46**, 337–339.
44. HOLLINGWORTH, H. L., Correlations of Achievement Within the Individual. *J. Exper. Psychol.*, 1925, **8**, 190–208.
45. HOLZINGER, K. J., *Statistical Methods for Students of Education.* Boston: Ginn & Co., 1928. Pp. 372.
46. HOTELLING, H., Analysis of a Complex of Statistical Variables into Principal Components. *J. Educ. Psychol.*, 1933, **24**, 417–441; 498–520.
47. HULL, C. L., The Meaningfulness of 320 Selected Nonsense Syllables. *Amer. J. Psychol.*, 1933, **45**, 730–734.

UNIVERSITY OF MICHIGAN LIBRARIES

48. HULL, C. L., The Conflicting Psychologies of Learning—A Way Out. *Psychol. Rev.*, 1935, **42**, 491–516.

49. HUNTER, W. S., Habit Interference in the White Rat and Human Subjects. *J. Comp. Psychol.*, 1922, **2**, 29–60.

50. HUNTER, W. S., A Reply to Professor Carr on "The Reliability of the Maze Experiment." *J. Comp. Psychol.*, 1926, **6**, 393–398.

51. HUNTER, W. S., Learning: IV. Experimental Studies of Learning. Pp. 497–570 in Murchison, C., *Handbook of General Experimental Psychology*. Worcester: Clark Univ. Press, 1934.

52. HUNTER, W. S., and RANDOLPH, V., Further Studies of the Maze with Rats and Humans. *J. Comp. Psychol.*, 1924, **4**, 431–445.

53. HUSBAND, R. W., A Comparison of Human Adults and White Rats in Maze Learning. *J. Comp. Psychol.*, 1929, **9**, 361–377.

54. HUSBAND, R. W., Analysis of Methods in Human Maze Learning. *J. Genet. Psychol.*, 1931, **39**, 258–278.

55. JACKSON, T. A., General Factors in Transfer of Training in the White Rat. *Genet. Psychol. Monog.*, 1932, **11**, 1–59.

56. JOHNSON, H. M., Some Follies of 'Emancipated' Psychology. *Psychol. Rev.*, 1932, **39**, 293–323.

57. KELLEY, T. L., Reliability of Test Scores. *J. Educ. Res.*, 1921, **31**, 370–379.

58. KELLEY, T. L., *Statistical Method*. New York: The Macmillan Co., 1924. Pp. 390.

59. KLINE, L. W., An Experimental Study of Associative Inhibition. *J. Exper. Psychol.*, 1921, **4**, 270–299.

60. KÖHLER, W., *Gestalt Psychology*. New York: H. Liveright, 1929.

61. KRUEGER, W. C. F., The Relative Difficulty of Nonsense Syllables. *J. Exper. Psychol.*, 1934, **17**, 145–153.

62. LANGDON, J. N., A Note on the Repetition of a Simple Motor Task. *Brit. J. Psychol.*, 1931, **22**, 55–61.

63. LANIER, L. H., Prediction of the Reliability of Mental Tests and Tests of Special Abilities. *J. Exper. Psychol.*, 1927, **10**, 69–113.

64. LEEPER, R., The Reliability and Validity of Maze Experiments with White Rats. *Genet. Psychol. Monog.*, 1932, **11**, 137–245.

65. LIGGETT, J. R., A Study of Maze Measures and of the Factors Involved in Maze Learning. *J. Genet. Psychol.*, 1930, **38**, 78–90.

66. LINDQUIST, E. F., and FOSTER, R. R., On the Determination of Reliability in Comparing the Final Mean-Scores of Matched Groups. *J. Educ. Psychol.*, 1929, **20**, 102–106.

67. LINDQUIST, E. F., A Further Note on the Significance of a Difference Between the Means of Matched Groups. *J. Educ. Psychol.*, 1933, **24**, 66–69.

68. LUH, C. W., The Conditions of Retention. *Psychol. Monog.*, 1922, **31**, No. 142. Pp. 87.

69. LUMLEY, F. H., An Investigation of the Responses Made in Learning a Multiple Choice Maze. *Psychol. Monog.*, 1931, **42**, No. 189. Pp. 61.

70. LUMLEY, F. H., Anticipation of Correct Responses as a Source of Error in the Learning of Serial Responses. *J. Exper. Psychol.*, 1932, **15**, 195–205.

71. LUMLEY, F. H., Anticipation as a Factor in Serial Maze Learning. *J. Exper. Psychol.*, 1932, **15**, 331–342.

72. LUMLEY, F. H., Anticipation and Erroneous Responses. *J. Exper. Psychol.*, 1934, **17**, 46–64.

73. MARTIN, P. R., and FERNBERGER, S. W., Improvement in Memory Span. *Amer. J. Psychol.*, 1929, **41**, 91–94.

74. MASLOW, A. H., The Effect of Varying Time Intervals Between Acts of Learning with a Note on Proactive Inhibition. *J. Exper. Psychol.*, 1934, **17**, 141–144.

75. MAURER, S., and CARR, H. A., II. The Empirical Determination of Maze Reliability. *J. Comp. Psychol.*, 1935, **20**, 291–308.

76. MCGEOCH, J. A., The Influence of Degree of Learning upon Retroactive Inhibition. *Amer. J. Psychol.*, 1929, **41**, 252–262.

77. MCGEOCH, J. A., The Influence of Associative Value upon the Difficulty of Nonsense-Syllable Lists. *J. Genet. Psychol.*, 1930, **37**, 421–430.

78. MCGEOCH, J. A., Forgetting and the Law of Disuse. *Psychol. Rev.*, 1932, **39**, 352–370.

79. MCGEOCH, J. A., The Influence of Degree of Interpolated Learning upon Retroactive Inhibition. *Amer. J. Psychol.*, 1932, **44**, 695–708.

80. MCGEOCH, J. A., Studies in Retroactive Inhibition: I. The Temporal Course of the Inhibitory Effects of Interpolated Learning. *J. Gener. Psychol.*, 1933, **9**, 24–43.

81. MCGEOCH, J. A., Studies in Retroactive Inhibition: II. Relationships Between Temporal Point of Interpolation, Length of Interval, and Amount of Retroactive Inhibition. *J. Gener. Psychol.*, 1933, **9**, 44–57.

82. MCGEOCH, J. A., Changes Accompanying Practice on Successive Samples of Verbal Material. *J. Gener. Psychol.*, 1933, **9**, 117–129.

83. MCGEOCH, J. A., The Vertical Dimensions of Mind. *Psychol. Rev.*, 1936, **43**, 107–129.

84. MCGEOCH, J. A., and OBERSCHELP, V. J., The Influence of Length of Problem and of Transfer upon Rational Learning and Its Retention. *J. Gener. Psychol.*, 1930, **4**, 154–170.

85. MCGINNIS, E., The Acquisition and Interference of Motor Habits in Young Children. *Genet. Psychol. Monog.*, 1929, **6**, 203–311.

86. MITCHELL, M. B., Anticipatory Place-Skipping Tendencies in the Memorization of Numbers. *Amer. J. Psychol.*, 1934, **46**, 80–91.

87. NYSWANDER, D. B., A Comparison of the High Relief Finger Maze and the Stylus Maze. *J. Gener. Psychol.*, 1929, **2**, 273–288.

88. PATERSON, D. G., ELLIOT, R. M., ANDERSON, L. D., TOOPS, H. A., and HEIDBREDER, E., Minnesota Mechanical Ability Tests. *Univ. of Minnesota*, 1930.

89. PAULSEN, G. B., The Reliability and Consistency of Differences in Motor Control. II. *J. Appl. Psychol.*, 1935, **19**, 166–179.

90. PEATMAN, J. G., and LOCKE, N. M., Studies in the Methodology of the Digit-Span Test. *Arch. of Psychol.*, 1934, **25**, No. 167. Pp. 35.

UNIVERSITY OF MICHIGAN LIBRARIES

91. PERRIN, F. A. C., On Experimental and Introspective Study of the Learning Process in the Maze. *Psychol. Monog.*, 1914, **16**, No. 70. Pp. 97.
92. PETERS, C. C., and VAN VOORHIS, W. R., A New Proof and Corrected Formulae for the Standard Error of a Mean and of a Standard Deviation. *J. Educ. Psychol.*, 1933, **24**, 620–633.
93. PETERSON, J., Thurstone's Measures of Variability in Learning. *Psychol. Bull.*, 1918, **15**, 452–456.
94. PETERSON, J., and ALLISON, L. W., Effect of Visual Exposure on the Rate and Reliability of Stylus-Maze Learning. *J. Gener. Psychol.*, 1930, **4**, 36–48.
95. PETERSON, J., and BARLOW, M. C., The Effects of Practice on Individual Differences. *The Twenty-Seventh Yearbook of The National Society for the Study of Education*, 1928, Part II, 211–230.
96. PETERSON, J., and LANIER, L. H., Studies in the Comparative Abilities of Whites and Negroes. *Ment. Meas. Monog.*, 1929, No. 5. Pp. 156.
97. PETERSON, J., and TELFORD, C. W., Results of Group and of Individual Tests Applied to the Practically Pure-Blood Negro Children of St. Helena Island. *J. Comp. Psychol.*, 1930–1931, **11**, 115–144.
98. RIETZ, H. L. (Ed.), *Handbook of Mathematical Statistics.* New York: The Houghton Mifflin Co., 1924. Pp. 221.
99. ROBINSON, E. S., Methods of Practice Equilibration. *Amer. J. Psychol.*, 1929, **41**, 153–156.
100. ROBINSON, E. S., *Association Theory To-Day.* New York: The Century Co., 1931. Pp. 142.
101. ROBINSON, E. S., and DARROW, C. W., Effect of Length of List upon Memory for Numbers. *Amer. J. Psychol.*, 1924, **35**, 235–243.
102. ROBINSON, E. S., and HERON, W. T., Results of Variations in Length of Memorized Material. *J. Exper. Psychol.*, 1922, **5**, 428–448.
103. SACKETT, R. S., The Influence of Symbolic Rehearsal upon the Retention of a Maze Habit. *J. Gener. Psychol.*, 1934, **10**, 376–398.
104. SAUER, F. M., The Relative Variability of Nonsense Syllables and Words. *J. Exper. Psychol.*, 1930, **13**, 235–246.
105. SCOTT, T. C., and NELSON, B. B., Factors Affecting the Reliability of the Maze: A Comparison of the High-Relief Finger Maze and an Improved Form of the Stylus Maze. *J. Gener. Psychol.*, 1932, **6**, 70–89.
106. SIIPOLA, E. M., and ISRAEL, H. E., Habit-Interference as Dependent upon Stage of Training. *Amer. J. Psychol.*, 1933, **45**, 205–227.
107. SKINNER, B. F., The Generic Nature of the Concept of Stimulus and Response. *J. Gener. Psychol.*, 1935, **12**, 40–65. (Bibliography.)
108. SPEARMAN, C., The Proof and Measurement of Association Between Two Things. *Amer. J. Psychol.*, 1904, **15**, 72–101.
109. SPEARMAN, C., Correlation Calculated from Faulty Data. *Brit. J. Psychol.*, 1909–1910, **3**, 271–295.
110. SPENCE, K. W., The Reliability of the Maze and Methods of Its Determination. *Comp. Psychol. Monog.*, 1932, **8**, No. 40. Pp. 45.
111. STEVENS, S. S., The Operational Basis of Psychology. *Amer. J. Psychol.*, 1935, **47**, 323–330.

112. Stevens, S. S., The Operational Definition of Psychological Concepts. *Psychol. Rev.*, 1935, **42**, 517–527.
113. Stone, C. P., The Reliability of Rat Learning Scores Obtained from a Modified Carr Maze. *J. Genet. Psychol.*, 1928, **35**, 507–521.
114. Stone, C. P., and Nyswander, D. B., The Reliability of Rat Learning Scores from the Multiple-T maze as Determined by Four Different Methods. *J. Genet. Psychol.*, 1927, **34**, 497–524.
115. Stroud, J. B., Lehman, A. F., and McCue, C., The Reliability of Nonsense-Syllable Scores. *J. Exper. Psychol.*, 1934, **17**, 294–304.
116. Thorndike, R. L., The Effect of Interval Between Test and Retest on the Constancy of the I.Q. *J. Educ. Psychol.*, 1933, **24**, 543–549.
117. Thorndike, R. L., Organization of Behavior in the Albino Rat. *Genet. Psychol. Monog.*, 1935, **17**, 1–70.
118. Thurstone, L. L., The Absolute Zero in Intelligence Measurement. *Psychol. Rev.*, 1928, **35**, 175–197.
119. Thurstone, L. L., *The Reliability and Validity of Tests.* Ann Arbor: Edwards Brothers, Inc., 1931. Pp. 113.
120. Thurstone, L. L., The Vectors of Mind. *Psychol. Rev.*, 1934, **41**, 1–32.
121. Thurstone, L. L., *The Vectors of Mind.* Chicago: Univ. of Chicago Press, 1935. Pp. 266.
122. Tolman, E. C., The Inheritance of Maze-Learning Ability in Rats. *J. Comp. Psychol.*, 1924, **4**, 1–18.
123. Tolman, E. C., and Nyswander, D. B., The Reliability and Validity of Maze-Measures for Rats. *J. Comp. Psychol.*, 1927, **7**, 425–460.
124. Tolman, E. C., Tryon, R. C., and Jeffress, L. A., A Self-Recording Maze with an Automatic Delivery Table. *Univ. Calif. Publ. Psychol.*, 1929, **4**, 99–112.
125. Tomilin, M. I., and Stone, C. P., Intercorrelations of Measures of Learning Ability in the Albino Rat. *J. Comp. Psychol.*, 1934, **17**, 73–88.
126. Tryon, R. C., Effect of the Unreliability of Measurement on the Difference Between Groups. *J. Comp. Psychol.*, 1926, **6**, 449–453.
127. Tryon, R. C., Demonstration of the Effect of Unreliability of Measurement on a Difference Between Groups. *J. Comp. Psychol.*, 1928, **8**, 1–22.
128. Tryon, R. C., Errors of Sampling and of Measurement as Affecting Difference Between Means. *J. Comp. Psychol.*, 1929, **9**, 191–195.
129. Tryon, R. C., The Reliability Coefficient as a Per Cent, with Application to the Correlation Between Abilities. *Psychol. Rev.*, 1930, **37**, 140–157.
130. Tryon, R. C., Studies in Individual Differences in Maze Ability. I. The Measurement of the Reliability of Individual Differences. *J. Comp. Psychol.*, 1930–1931, **11**, 145–170.
131. Tryon, R. C., Individual Differences in Maze Ability. II. The Determination of Individual Differences by Age, Weight, Sex, and Pigmentation. *J. Comp. Psychol.*, 1931, **12**, 1–22.
132. Tryon, R. C., Studies in Individual Differences in Maze Ability. III. The Community of Function Between Two Maze Abilities. *J. Comp. Psychol.*, 1931, **12**, 95–115.

UNIVERSITY OF MICHIGAN LIBRARIES

133. TSAI, C., A Comparative Study of Retention Curves for Motor Habits. *Comp. Psychol. Monog.*, 1924, **2**, No. 11. Pp. 29.

134. VALENTINE, W. L., and MEYER, M., A Description of the Lectometer and the Reliability of Lectometer Scores. *J. Gener. Psychol.*, 1930, **4**, 407–415.

135. WALKER, H. M., Concerning the Standard Error of a Difference. *J. Educ. Psychol.*, 1929, **20**, 53–60.

136. WALKER, H. M., *Studies in the History of Statistical Method.* Baltimore: The Williams and Wilkins Co., 1929. Pp. 229.

137. WARD, L. B., *Retention over Short Intervals of Time.* Ph.D. Thesis, Unpublished, Yale University, 1934.

138. WARDEN, C. J., The Relative Economy of Various Modes of Attack in the Mastery of a Stylus Maze. *J. Exper. Psychol.*, 1924, **7**, 243–275.

139. WILLOUGHBY, R. R., The Concept of Reliability. *Psychol. Rev.*, 1935, **42**, 153–165.

140. WILTON, J. R., The Effects of Observational Errors and Other Factors upon Correlation Coefficients in Psychology. *J. Exper. Ped.*, 1914, **2**, 301–304.

141. WITMER, L. R., The Association Value of Three-Place Consonant Syllables. *J. Genet. Psychol.*, 1935, **47**, 337–360.

142. WOODROW, H., The Effect of Type of Training upon Transference. *J. Educ. Psychol.*, 1927, **18**, 159–172.

143. WOODROW, H., Quotidian Variability. *Psychol. Rev.*, 1932, **39**, 245–256.

144. WOODYARD, E., *The Effect of Time upon Variability.* New York: Bureau of Publications, Teachers College, Columbia University, 1926. Pp. 56.

145. WYATT, S., The Quantitative Investigation of Higher Mental Processes. *Brit. J. Psychol.*, 1913, **6**, 109–135.

146. YULE, G. U., *On Introduction to the Theory of Statistics.* (Ninth Ed.) London: Chas. Griffin and Co., Ltd., 1929. Pp. 424.

SPECIAL REVIEW

THURSTONE'S VECTORS OF MIND[1]

BY HENRY E. GARRETT
Columbia University

The contribution which this book makes to the field of mental measurement can best be appreciated, perhaps, in the light of the historical background out of which it comes. About thirty years ago, Professor Charles Spearman proposed his two-factor theory, which states that an individual's performance upon a mental test can best be understood in terms of two factors *g* and *s*. The general factor, *g*, may be identified somewhat loosely with general intelligence or general level, and is conceived of as accounting for the intercorrelations among mental tests. The specific factor, *s*, is peculiar to each test, and together with *g* is thought of as determining a subject's score. A test score may be high because of much *g*, or much *s*, or considerable amounts of both.

Except for criticism by Professor E. L. Thorndike, who championed the view that numerous factors rather than two contribute to ability, for twenty years or so Spearman's theory met with little favor or disfavor in this country. In England, the theory of two-factors won many adherents, and at least one able and persistent critic in Godfrey Thomson, who has attacked both its mathematical foundations and its psychological interpretation. Thomson upholds a " sampling theory " which accounts for mental test correlations by assuming the existence of many factors which combine in various ways and numbers.

The beginning of real interest in " factor analysis " in this country may be said to date from the publication in 1928 of T. L. Kelley's *Crossroads in the Mind of Man.* Since this time, several able workers have examined the mathematics underlying Spearman's theory, and have conducted extensive experiments in the field of mental organization. Among the most active workers in this field has been Professor L. L. Thurstone. His first paper on multiple

[1] Thurstone, L. L., *The Vectors of Mind.* Chicago: The University of Chicago Press, 1935. Pp. xv+266.

UNIVERSITY OF MICHIGAN LIBRARIES

factor analysis was published in 1931. This was followed by several studies in which he amplified and extended Spearman's methods and contributed new techniques of his own to the solution of the multiple factor problem. *The Vectors of Mind* summarizes Thurstone's multiple factor methods to date, and presents much new material not hitherto published.

Thurstone performs a real service to students by opening his book with a "Mathematical Introduction" in which are presented the elements of matrix theory. The theory of matrices as such is relatively new even in mathematics (it was introduced about 1843) and the majority of psychologists, I suspect, are unfamiliar with it, except, perhaps, for a nodding acquaintance with determinants. I am sure that no one who has not worked through Thurstone's "Introduction" or equivalent material can possibly follow his arguments or understand his techniques. But I am doubtful whether psychologists who possess little mathematical aptitude can master this material, unless they are willing to spend more time on it than most of them are willing—or able—to devote. The reader who finds the "Introduction" hard going may be assured (if it is any comfort) that Thurstone's discussion of matrix algebra is far more lucid than is the treatment of this topic in current mathematical textbooks.

Chapters I and II deal with "The Factor Problem" and "The Fundamental Factor Theorem," respectively. A few definitions at the outset will serve to clarify the discussion here. Thurstone defines a *trait* as *any* attribute of an individual. Traits are differentiated into those "which are descriptive of the individual as he appears to others, and those (traits) which are exemplified primarily in things he can do" (p. 48). *Abilities* are traits of the last kind ("things he can do"); *tests* define abilities through *scores*. The total variance (σ^2) of a test is the sum of the squares of all factors, common, specific, and error, which the test contains and is equal to *1* when the scores in the different factors are expressed in σ-units; the reliability of a test is that part of the variance attributable to common and specific factors; the communality is that part of the variance due to common factors; and specificity that part of the variance due to specific factors. The uniqueness of a test is that part of the variance due to specific and error factors.

The object of all factor analysis is to discover independent reference values (psychologically these may be later identified as "primary" traits) which will serve to reproduce a table of correlations. It is well known, however, that a given set of correlations

may be reproduced by a large number of "factor patterns." Hence it becomes necessary to lay down certain conditions or to make certain postulates which will provide for a mathematically unique factor solution, as well as one which is, in some sense, psychologically better than alternative patterns. Thurstone lays down the fundamental factor theorem that the number of independent factors required to reproduce the intercorrelations of n tests is equal to the rank of the correlational matrix, *i.e.* the table of intercorrelations. (For meaning of "rank" of a matrix, see p. 10.) The rank of a correlational matrix is not its apparent rank, which is usually n when there are n tests, but its minimum effective rank—its rank when errors of sampling and errors of measurement are eliminated. To illustrate, the rank of a correlational matrix when there is only one common factor (Spearman's g, say) is *1;* and hence all of the determinants of order two or above should equal zero. But the tetrad differences (two rowed determinants) are rarely zero *exactly,* though they may deviate insignificantly from zero. Therefore, the minimum effective rank of a Spearman matrix is *1,* although its apparent rank may be higher. The object of factor analysis, then, is to find experimentally the minimum effective rank of a matrix of intercorrelations.

Stated in terms of matrix algebra, the factor problem resolves itself into the search for a factor matrix (F) which when multiplied by its transpose (F′) will reproduce the reduced correlational matrix R_0. By reduced correlational matrix, Thurstone means the matrix of intercorrelations in which communalities have been entered in the main diagonal; but as will appear later in this review, this does not seem to be a crucial requirement. Thurstone gives (p. 75) a simple test by means of which one can determine the number of independent factors which one may expect to find in n tests. To identify, for example, three independent factors, one needs at least six tests.

Given a unique pattern (or factorial matrix) an individual's score in a given test may be described by the weighted linear equation,

$$s_{j1}=a_{j1}x_{11}+a_{j2}x_{21}+a_{j3}x_{31}+\ \ldots\ +a_{jp}x_{q1}$$

in which s_{j1} represents the standard score of individual i in test j; the x's represent the subject's scores in q independent factors; and the a's are the weights of the different factors, *i.e.* the degree to which each factor enters into the given test score. An individual's score, therefore, depends upon (1) the extent to which he possesses the given factors (the x's) and the weight or importance of each

factor in the given test-ability. In Spearman's two-factor theory, the above equation becomes

$$s_{j1} = a_{j1}\, g_1 + a_{j2}\, s_1$$

Both of these equations, whether for two or more factors, depend for their validity upon Taylor's expansion, whereby almost any function no matter what the basic relationships involved (*e.g.* multiplicative, logarithmic, etc.) can be expressed to close approximation by the sum of a set of terms.

Chapter III, "The Centroid Method," and Chapter IV, "The Method of Principal Axes," deal with the fundamental methods of extracting independent factors from a given set of intercorrelations. Chapter V, "The Special Case of Rank One," outlines a method of calculating the common factor loadings when there is only one factor present, *i.e.* when the rank of the correlational matrix is *1*. The method given by Thurstone is relatively easy to apply and requires less calculation than does the method of tetrad-differences. The centroid or center of gravity method is Thurstone's *Simplified Multiple Factor Method* published in 1933. As given here, it has certain refinements, especially as regards sign reversals in computing the factor loadings of tests "reflected" through the origin. The method of principal axes will be found in a mathematically less general, but probably more comprehensible form, in *Theory of Multiple Factors,* published in 1933.

The principles of the centroid method may be best understood, perhaps, through a geometrical description or picture of the relations which it assumes between factors and tests. Suppose that one conceives of his tests as dots upon the surface of a sphere, and of the factors as axes of the sphere. Then the correlation between any two tests will equal the cosine of the angle between the lines (test vectors) joining the dots to the center of the sphere; and the correlation of a test with a factor is the projection of the test vector upon the axis representing the factor (reference vector). Since the cosine of an angle increases as the size of its angle decreases, those tests which are highly correlated will appear close together (*e.g.* in the form of a cluster) upon the surface of the sphere.[2]

The coördinates of the centroid or center of gravity of our system of points or test dots are the means of the projections of the n tests upon the reference vectors, taken in order. The centroid, therefore,

[2] Strictly speaking, our test dots lie *below* and not *upon* the surface of the sphere; only when the communality is *1*, *i.e.* when the test contains no specific factors or chance errors, does a test dot lie upon the surface of the sphere.

will lie somewhere between the origin of the sphere and the main cluster of test dots. In order to extract the first factor, the sphere is rotated so that the reference axis passes through the centroid. The projections of the test vectors upon this axis or reference vector give the first factor loadings; and these loadings make the largest contribution to the variance of the test battery of any of the centroid factors. After the first factor is extracted, the residual correlational matrix is investigated for a second factor. The centroid now lies at the origin, since all of its coördinates are zero except that point which determined the centroid's position on the first axis. This axis is removed with the first factor. Hence, the method employed in extracting the first factor is not directly applicable to the calculation of a second factor. Thurstone has devised an ingenious scheme for finding the second factor. This consists in "reflecting" (by changing signs) a test through the origin to the opposite end of a diameter when by so doing a new cluster of dots (tests) may be built up and a new centroid located. After reflection, the sphere is again rotated and a second axis passed through the new centroid. After the second factor is extracted, the factor loadings are given their original (unchanged) signs. The process of reflecting tests and extracting new factors is continued until the residual correlation is approximately zero.

The principal axes in the method of that name are those axes of our hypothetical sphere upon which the projections of the test vectors are maximal. The method of principal axes as described by Thurstone is essentially the method of principal components devised by Hotelling, and described by him in the *Journal of Educational Psychology* in 1933.[3] Thurstone discards the principal axes in favor of the centroid method because in the former method "(1) the number of factors is a function of the number of tests in the battery, and (2) about half of the factor loadings beyond the first are necessarily negative" (p. 120). He rejects Hotelling's principal components on the ground that the placing of *1*'s in the main diagonal of the correlational matrix (*i.e.* using correlations corrected for attenuation, and unity for reliability coefficients) implies that the total variance of each trait or test can be described by common factors despite the fact that the specific factors remain even when chance errors are eliminated. Hotelling's method extracts as many factors as there are tests, which procedure, Thurstone argues, excludes the possi-

[3] Hotelling, A., Analysis of a Complex of Statistical Variables into Principal Components. *J. Educ. Psychol.*, 1933, **24**, 417–441; 498–520.

bility of unique variance arising from sampling and chance and specific factors. More generally, Hotelling's method is criticized because, according to Thurstone, (1) to postulate as many factors as there are tests does not provide a scientifically useful solution; and (2) because the factor loadings of a given test are a function of the particular battery of which it is a member, and hence can have no stability and no precise psychological meaning.

I do not think that these criticisms of the method of principal components are well taken, except, perhaps, the last, and that is true only in those special cases wherein a battery contains a large number of tests of one sort (*e.g.* " verbal " tests) and only one or two tests of a distinctly different kind, say, " number " tests. An overwhelming " verbal factor " might distort the factorial description of number or spatial tests; but this is not true when the battery is large, and when it samples a number of different abilities. The factorial descriptions given by the centroid and principal axes methods of Brigham's 15 tests, for example, are almost identical both in size of factor weights and in sign attached, as Thurstone's own analysis shows (pp. 131–132). Hotelling's method of principal components allows the use of reliability coefficients in the main diagonal of the correlation table, as well as of *1*'s, so that Thurstone's criticism applies only to one variation of the method. Moreover, Hotelling's method is an iterative one, in which the weights of the successive factors, that is, their contributions to the total variance of the test battery, rapidly get smaller, so that it is rarely necessary to calculate more than three or four factors no matter how large the test battery. When all components are computed, reliabilities being placed in the main diagonal, the last few become effectively " specifics," as one may readily discover by applying the method. Thurstone's use of communalities in the main diagonal seems to me to be less defensible than Hotelling's use of *1*'s. In the first place, the true communality of a test is unknown, as Thurstone admits. It must lie somewhere between the reliability coefficient of the test and the highest correlation of the given test with a member of the battery. Thurstone arbitrarily takes the largest correlation coefficient of the test with another test of the battery as the best estimate of its communality. However, these estimated communalities will certainly vary widely as the test is moved from one battery to another, or if the battery itself is lengthened or shortened. Reliability coefficients would seem, therefore, to be far more stable entries for the main diagonal than estimated communalities.

A crucial question which arises in all factor analysis is that of whether "body" or psychological meaning can be given to the factors extracted from a table of intercorrelations. The usefulness of factor analysis in the study of mental organization hinges upon whether an affirmative answer can be given to this question. Are the factors isolated by analytic methods simply and solely mathematical entities, or glorified fractions which are hypostatized into mental "faculties"; or can they be conceived of as representing the operative strength of true abilities? It has been said, and with much justification, that since the factors extracted from a correlational table are simply averages based upon the variables concerned, they must of necessity partake in varying degree of all of the aptitudes which conceivably condition performance upon the tests of the battery. Hence, if subjects differ in age, sex, and educational background, and if the tests of the battery are numerous, a factor may well be a kind of "psychological hash" comprising odds and ends of all sorts.

Thurstone attacks the problem of the identification of factors in Chapter VI, "Primary Traits," and Chapter VIII, "Isolation of Primary Factors." Certain criteria are set up for a "primary trait" which can best be understood, I think, by a geometrical description of our tests and factors in terms of an n-dimensional sphere. The tests in a battery, represented by dots on the surface of the sphere, may be scattered indiscriminately over the surface area. Very often, however, these test dots exhibit a definite arrangement which gives a strong presumption of underlying order. Suppose, for example, that the tests in a battery fall into three well defined groups; and that each group falls along or close to the circumference of a great circle. A configuration of tests of this sort is said to exhibit *simple structure;* if the three intersections of the planes determined by the tests are perpendicular to each other, it exhibits *orthogonal simple structure;* if the intersections of the planes are not perpendicular, it exhibits *oblique simple structure.* When structure is found in a battery of tests, Thurstone calls the axes of intersection of the determining planes *primary vectors.* When primary vectors can be given psychological meaning, they become *primary traits.* A primary vector or axis has substantially zero correlation with all of the tests not lying in the planes which determine it. Therefore, a primary trait or primary factor will reduce the number of factors per test which will serve to account for the intercorrelations in the table. When primary traits are present, they define a factor "set-up" which

has a definite and unique organization. Primary traits appear only when the test and reference axis configuration shows structure; hence the inference is strong that such structure is indicative of some underlying pattern within the abilities concerned.

Thurstone outlines five methods by means of which one can test for a primary trait. Of these, the method of oblique axes and the method of averages seem to me to make fewer assumptions and to be most useful. If in a test battery one can establish the existence of "verbalness" as a primary trait, he may be assured of an underlying arrangement prevailing among his tests. I am not sure that a geometrically "pure" trait necessarily implies a psychologically pure counterpart. But it does seem that definite structure renders highly improbable the hypothesis that a given factor is simply a hodge-podge—the average of many heterogeneous abilities. The identification of primary traits through the discovery of structure is extremely ingenious, and may, perhaps, be the most valuable part of this book. It should certainly be followed up experimentally.

Chapter VIII, "The Positive Manifold," attacks another well known stumbling block in factor analysis, namely, that of explaining negative factor loadings. In analyzing a correlational matrix of personality tests, it is not hard to conceive of a factor, identified with "submission," say, being negatively correlated with tests of "extroversion." Also, one can conceive of a factor of "motor dexterity" having zero or negative weight in tests of abstract reasoning. But it is hard to explain the large number of negative factor loadings which one gets in a factorial description of the usual battery of mental tests. "Memory," "mental speed," or "number ability" may perhaps have zero weight in certain abilities, but would they ever actually be deterrents? When the intercorrelations of a given test battery are mainly positive or zero, as is true of mental tests generally, the matter of negative factors offers no especial difficulties. A primary trait, if one exists, has positive correlations of necessity with the test group in which it is present, and zero or near zero correlations with other tests of the battery. When there are many negative correlations, the problem of finding traits which are confined to the positive section of the sphere or to the positive manifold becomes more difficult. Several approaches to a solution of this problem are offered by Thurstone (pp. 202–205).

The special case in which factors are *unitary*—either present or absent as are presumably genetic elements—is also considered in Chapter VIII. If the unitary elements are equally weighted as

regards their contribution to the variance of the traits which they determine, then the correlation between two tests j and k reduces to the well-known formula

$$r_{jk}=\frac{n_{jk}}{\sqrt{n_j\, n_k}}$$

in which n_{jk} equals the number of common elements in j and k, and n_j and n_k are the total number of elements in j and k, respectively. If these elements represent genetic factors they must necessarily be integral, so that the correlation of two traits which are genetically determined will vary directly with the number of elements which they possess in common and the number of elements in each. Thurstone offers a type of analysis by which knowing the number of elements within each of two traits (the complexity of the traits) one may infer from the correlation the number of unitary elements which they possess in common. Analysis of this sort offers many interesting possibilities.

Chapter IX, "Orthogonal Transformations," gives a method of investigating a factor matrix (set of calculated factor weights) for primary traits when the axes representing the primary traits are orthogonal, *i.e.* perpendicular to each other. The method is not especially difficult to follow, if one has mastered the matrix algebra of the "Mathematical Introduction." The book closes with Chapter X, on the "Appraisal of Abilities." In this chapter, a regression equation is derived by means of which the "score" of an individual in a primary trait may be estimated from his scores on the tests of the battery. An appendix outlines in detail the steps to be followed in calculating independent factors by the centroid method.

It is difficult to evaluate the psychological value of a book which like this one is almost entirely mathematical. Thurstone writes in his preface that "The future development of factor analysis will probably require more mathematical competence than we (psychologists) can supply in our own ranks." This statement, I suspect, will bring strong dissent from many psychologists who already believe that "psychometrics" stresses "metrics" to the almost total eclipse of the psychology involved. Whether Thurstone is right or not in his call for more and better mathematics in psychology remains for the future to decide. A factorial description of mental tests offers a precise analysis the truth of which can be *checked* by experiment. Such an analysis has a marked disadvantage, of course, over descriptions in terms of so-called psychological components, con-

cepts, dispositions, and the like, since the latter, being unverifiable, are always true. This is a dubious honor, however. To say that human behavior is exceedingly complex and is conditioned by many factors, both environmental and hereditary, is doubtless true, but is hardly valuable. Relations, to be of scientific value, must be quantitative whether one is dealing with the attraction of bodies or with the formation of conditioned reflexes.

The Vectors of Mind is an important addition to the literature on mental measurement. Its contribution, however, is almost entirely methodological; and its usefulness to the psychology of mental organization is still a promise for the future. In order that this promise be realized, the thing most needed now is for these new methods to be applied experimentally. When and if primary traits are isolated, and their reality verified in terms of acceptable criteria, we shall be on the way toward a scientific description of human nature.

BOOKS RECEIVED

BLUMENFELD, VON W., *Jugend als Konfliktsituation.* Berlin: Philo Verlag G.m.b.H., 1936. Pp. 122.

CAMPBELL, N. M., *The Elementary School Teacher's Treatment of Classroom Behavior Problems.* T. C. Contr. to Educ. No. 668. New York: Bureau of Publications, Teachers College, Columbia University, 1935. Pp. vi+71.

COLLINS, M., and DREVER, J., *Psychology and Practical Life.* London: University of London Press, 10 Warwick Lane, E.C.4, 1936. Pp. viii+307.

DREVER, J., and COLLINS, M., *Performance Tests of Intelligence: A Series of Non-linguistic Tests for Deaf and Normal Children.* (Second Edition.) Edinburgh: Oliver and Boyd, Tweeddale Court, 1936. Pp. 56.

DURET, R., *Les aspects de l'image visuelle.* Paris: Boivin et Cie, 5 Rue Palatine, 1936. Pp. vi+148.

FOXE, A. N., *Crime and Sexual Development.* Glens Falls, New York: The Monograph Editions, 1936. Pp. 91.

JANKÉLÉVITCH, V., *L'ironie.* Paris: Félix Alcan, 108 Boulevard St.-Germain, 1936. Pp. 149.

KUGELMASS, I. N., *Growing Superior Children.* New York: D. Appleton-Century Company, 1935. Pp. xvi+568.

LENTZ, T. F., *The C-R Opinionaire. A Measure of Conservatism and Radicalism for College Students and Adult Groups.* St. Louis: Character Research Institute, Washington University, 1935.

MELVIN, A. G., *The Activity Program.* New York: John Day and Reynal and Hitchcock, 1936. Pp. ix+275.

MOLONY, W. O'S., *New Armor for Old.* New York: Henry Holt and Company, 1935. Pp. 442.

RUSU, L., *Le sens de l'existence dans la poésie populaire roumaine.* Paris: Félix Alcan, 108 Boulevard St.-Germain, 1935. Pp. 123.

SHAFFER, L. F., *The Psychology of Adjustment: An Objective Approach to Mental Hygiene.* Boston: Houghton Mifflin Company, 1936. Pp. xix+600.

SOLLIER, P., and DRABS, J., *La psychotechnique.* Bruxelles: Comité Central Industriel de Belgique, 33 Rue Ducale; and Paris: Félix Alcan, 108 Boulevard St.-Germain, 1935. Pp. xviii+189.

VAUGHAN, W. F., *General Psychology.* Garden City, New York: Doubleday, Doran and Company, 1936. Pp. xxi+634.

WEISENBURG, T., ROE, A., and MCBRIDE, K. E., *Adult Intelligence.* New York: The Commonwealth Fund, 1936. Pp. xiii+155.

NOTES AND NEWS

DR. LEONARD CARMICHAEL, director of the psychological laboratory and laboratory of sensory physiology at Brown University, has accepted a position as chairman of the department of psychology and director of the psychological laboratory at the University of Rochester. Dr. Karl U. Smith, instructor in psychology at Brown, is also resigning at Brown to accept an instructorship at Rochester.

By arrangement between the administrations at Brown and Rochester, research apparatus and graduate students working directly with Dr. Carmichael and Dr. Smith are to be transferred to the new research laboratory of psychology at Rochester which is being established. It is planned to develop the new psychological laboratory at Rochester in close collaboration with the other scientific research departments of the University and especially with the departments of neurology and physiology in the Medical School and with the Institute of Optics of the University.

Besides being chairman of the department of psychology, Dr. Carmichael is also to be dean of the Faculty of Arts and Sciences at Rochester.

DR. WALTER S. HUNTER, G. Stanley Hall Professor of Genetic Psychology at Clark University since 1925, has resigned to accept a professorship of psychology at Brown University.

DR. CLARENCE H. GRAHAM, assistant professor of psychology at Clark University since 1932, has been appointed assistant professor of psychology at Brown University.

DR. J. MCVICKER HUNT, PH.D. Cornell University, 1933, and National Research Council Fellow 1933–1935, has been appointed instructor in psychology at Brown University.

DR. C. LLOYD MORGAN, professor emeritus in the University of Bristol, who was the first vice-chancellor of the University, died on March 6 at the age of 84 years. Dr. Morgan had filled the chair of geology and biology at University College, Bristol, from 1883 to 1887, when he became principal. He was appointed chancellor in 1910. This post he relinquished after a few months and was then appointed to fill the new chair of psychology and ethics. This chair he held until his retirement in 1919.—From *Science.*

THE psychologists of the State of Oregon held their first meeting at the University of Oregon, February 28 and 29, under the chairmanship of Professor Howard R. Taylor. Two sessions were held, one devoted to the topic, *The Teaching of Elementary Psychology,* and the other to a discussion of research projects. Friday evening after an informal dinner, Dr. Arnold Gesell's sound film *Life Begins* was presented. Professor William Griffith of Reed College will be chairman of the meeting next year, which is to be held at Reed College. Dr. Calvin S. Hall of the University of Oregon was elected secretary.

THE COMMITTEE FOR THE STUDY OF SUICIDE, INC., was incorporated last December under the laws of the State of New York, and began its activities early in January. It plans to undertake a comprehensive study of suicide as a social and psychological phenomenon. Dr. Gerald R. Jameison is president of the committee, Mr. Marshall Field, vice-president, Dr. H. A. Riley, treasurer, and Dr. Gregory Zilboorg, secretary and director of research. Dr. Henry E. Sigerist, professor of the history of medicine at Johns Hopkins University, and Dr. Edward Sapir, professor of anthropology at Yale University, are consultant members of the committee. The executive offices are located at Room 1404, the Medical Arts Center, 57 West 57th Street, New York City.